VBS VVH All-Hadronic

Internal status report

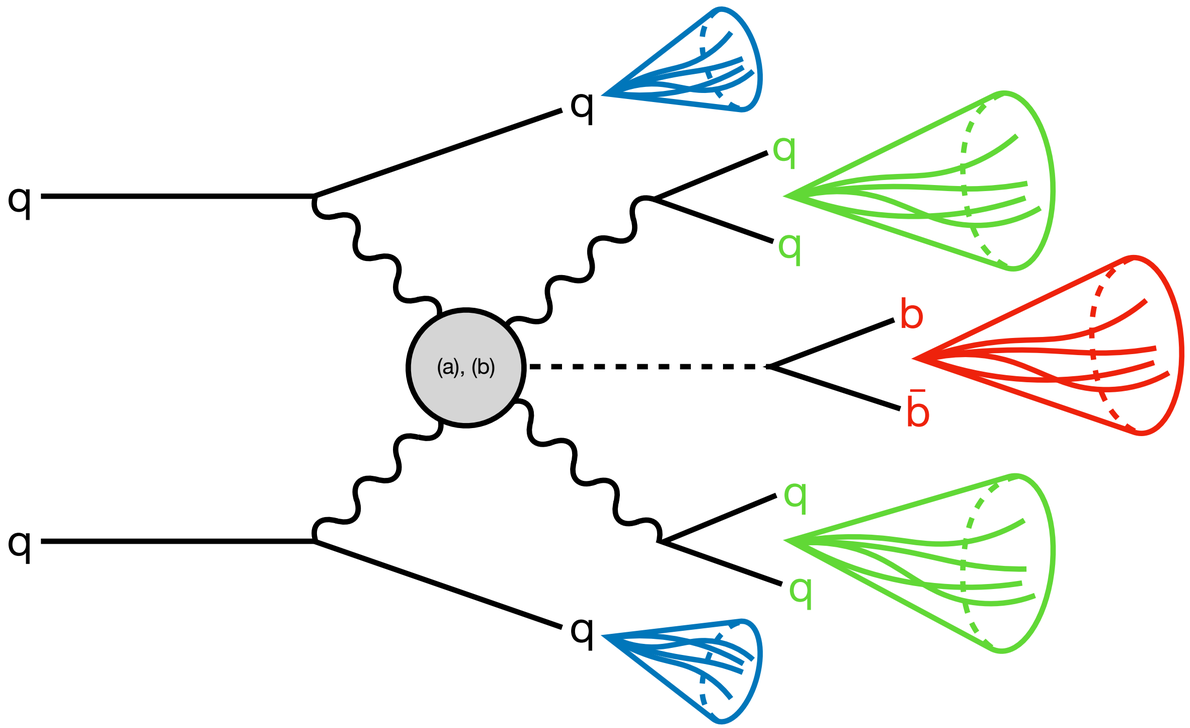
May 26th, 2023

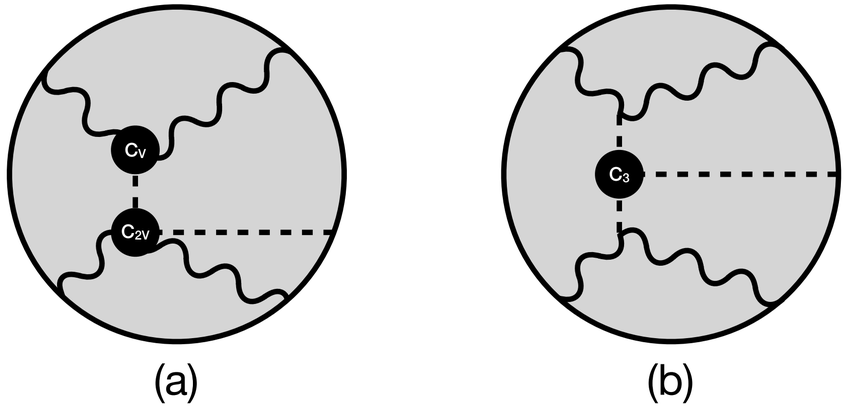
*P. Chang, L. Giannini, **J. Guiang**, Y. Xiang, E. Zenhom*

Analysis Overview

All-Hadronic VBS VVH

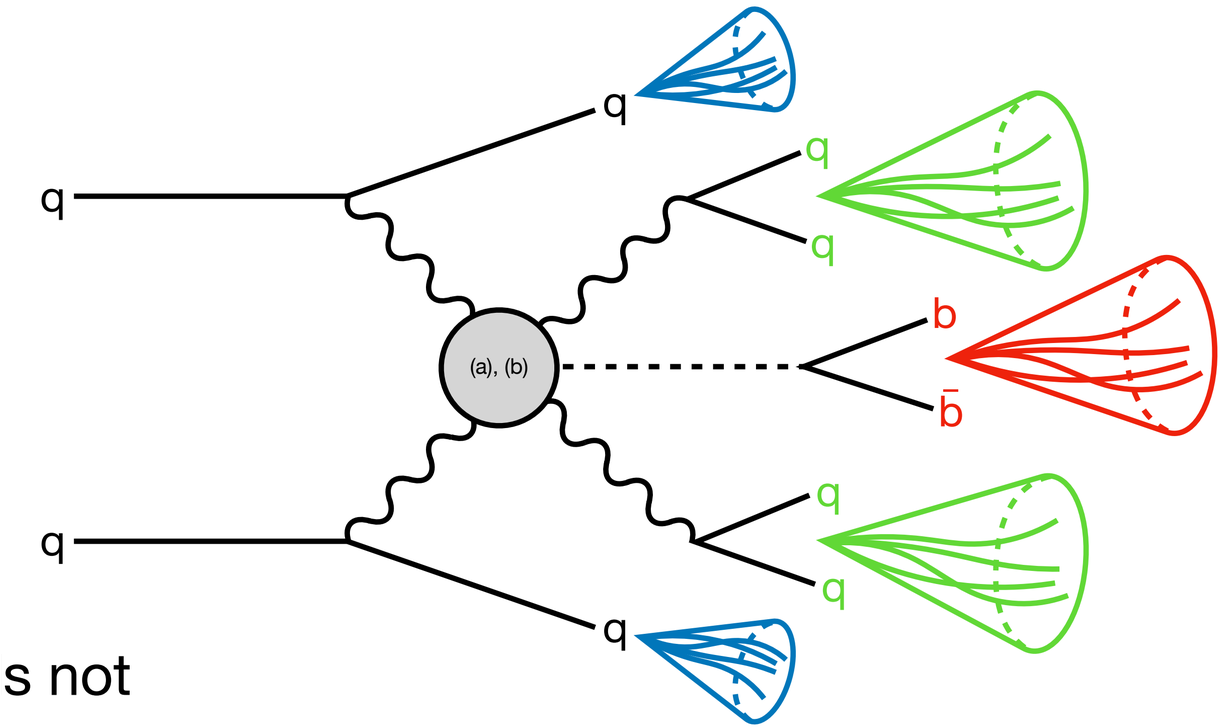
- Targeting the following final states:
  - VBS WWH → qq qq qq $b\bar{b}$
  - VBS ZZH → qq qq qq $b\bar{b}$
  - VBS WZH → qq qq qq $b\bar{b}$
- Sensitive to C_{2V}, C_3, and C_V in principle
- BSM signature:
  - W/Z/H jets with large p_T
  - VBS jets with large $\Delta\eta_{jj}$, M_{jj}

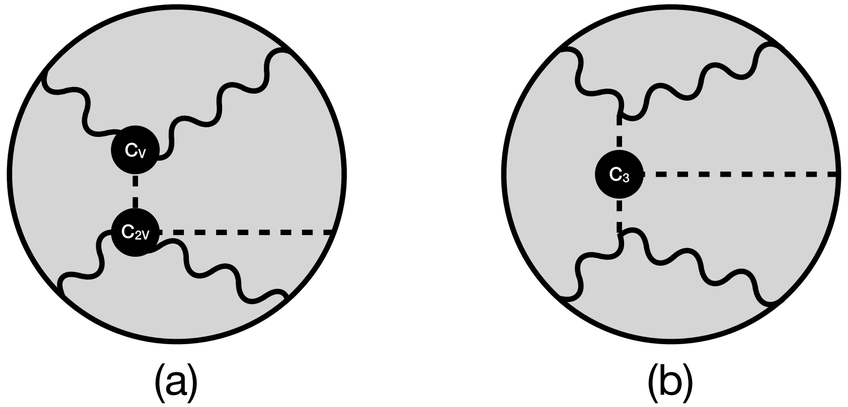




All-Hadronic VBS VVH

- One interesting N_{jets} vs. $N_{fatjets}$ channel:
  - **≥ 3 AK8 fat jets, ≥ 2 AK4 jets** (right)
  - ~~2 AK8 fat jets, ≥ 4 AK4 jets~~
  - ~~2 AK8 fat jets, 3 AK4 jets~~
- From previous studies, $N_{fatjets} < 3$ channels not worthwhile pursuing right now





Skim + Triggers + 3 Fat Jet Region

Yields scaled to lumi×σ, rounded for readability

Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%

Object	Skim Selection
Leptons (μ, e)	== 0 veto*
Fat Jets	≥ 2 AK8 jets w/ p_T > 300 GeV AND |η| < 2.5 AND mass > 50 GeV AND M_{SD} > 40 GeV AND fat jet ID > 0
Jets	≥ 2 AK4 jets w/ p_T > 20 GeV AND passes tight jet ID AND ΔR(jet, fat jet) > 0.8

*Using the ttH lepton ID

Year	HLT path
2016	HLT_PFHT800 HLT_PFHT900 HLT_AK8PFHT650_TrimR0p1PT0p03Mass50 HLT_AK8PFHT700_TrimR0p1PT0p03Mass50 HLT_AK8PFJet450 HLT_AK8PFJet360_TrimMass30 HLT_AK8DiPFJet280_200_TrimMass30 HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV_p20
2017	HLT_PFHT1050 HLT_AK8PFHT800_TrimMass50 HLT_PFJet320 HLT_PFJet500 HLT_AK8PFJet320 HLT_AK8PFJet500 HLT_AK8PFJet400_TrimMass30 HLT_AK8PFJet420_TrimMass30
2018	HLT_PFHT1050 HLT_AK8PFHT800_TrimMass50 HLT_PFJet500 HLT_AK8PFJet500 HLT_AK8PFJet400_TrimMass30 HLT_AK8PFJet420_TrimMass30

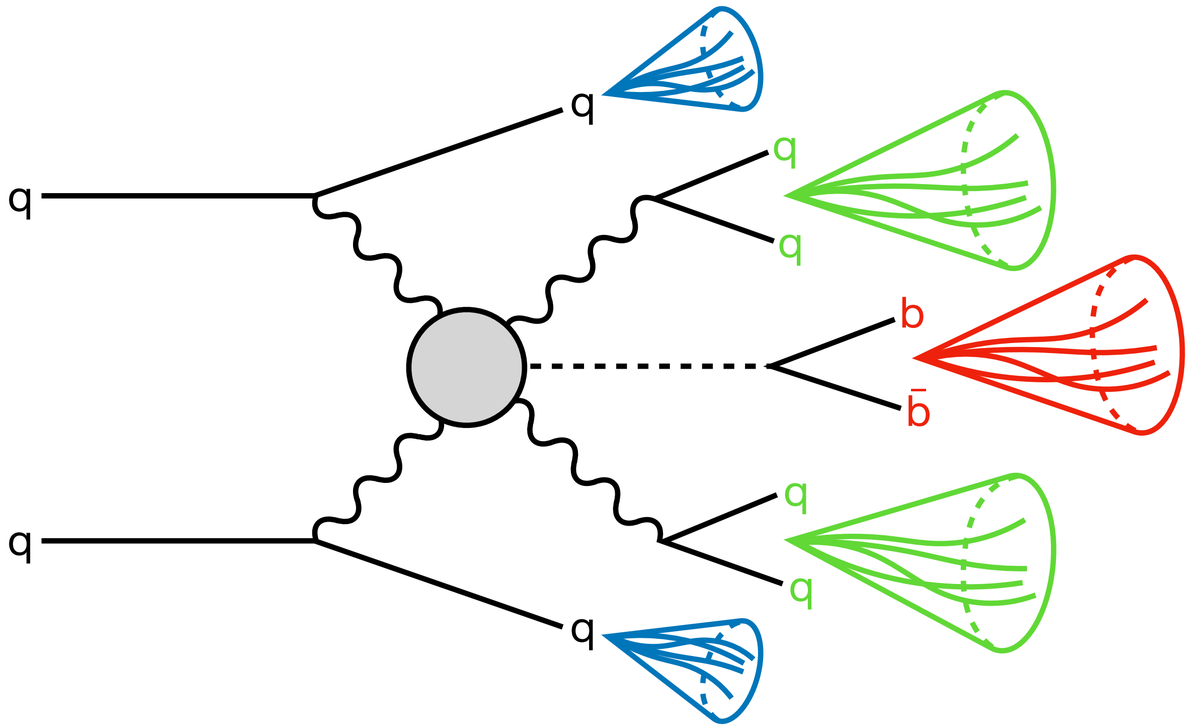
Taken from B2G-21-003

Object Selection

Yields scaled to lumi×σ, rounded for readability

Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%

Object	Selections
AK8 jets	• Same as skim • max(p_T) > 550 GeV (HLT plateau)
H→$b\bar{b}$ fat jet	• Has max(ParticleNet Xbb)
V→qq fat jets	• Not the H→$b\bar{b}$ candidate • Leading and next-leading in p_T
AK4 jets	• Same as skim
VBS (AK4) jets	• p_T > 30 GeV • For > 2 candidates: • Take pair with maximum $\lvert\Delta\eta_{jj}\rvert$



UC San Diego

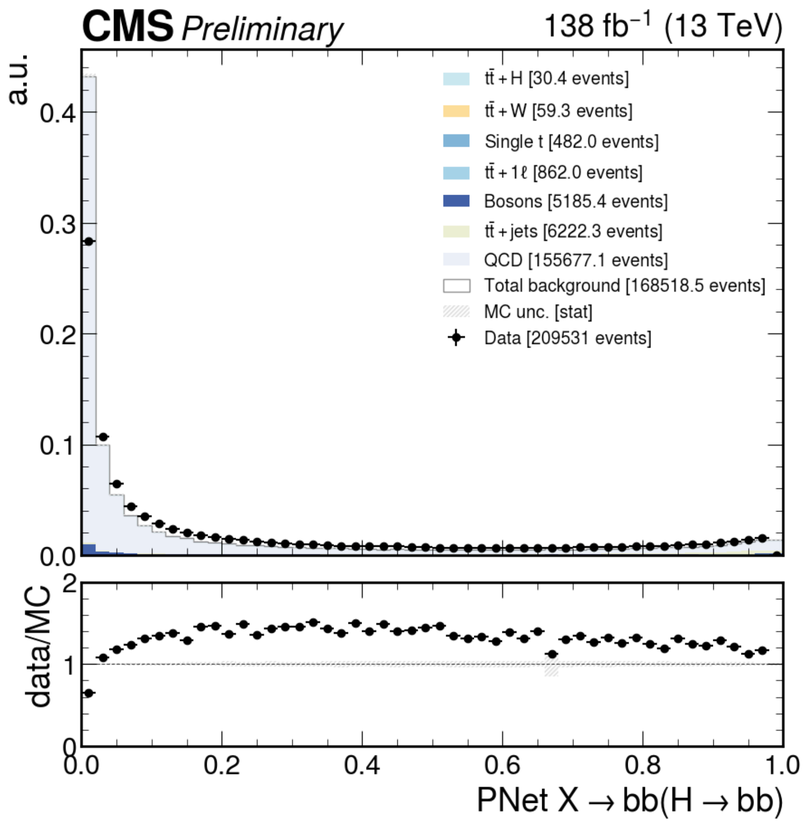
Object Selection

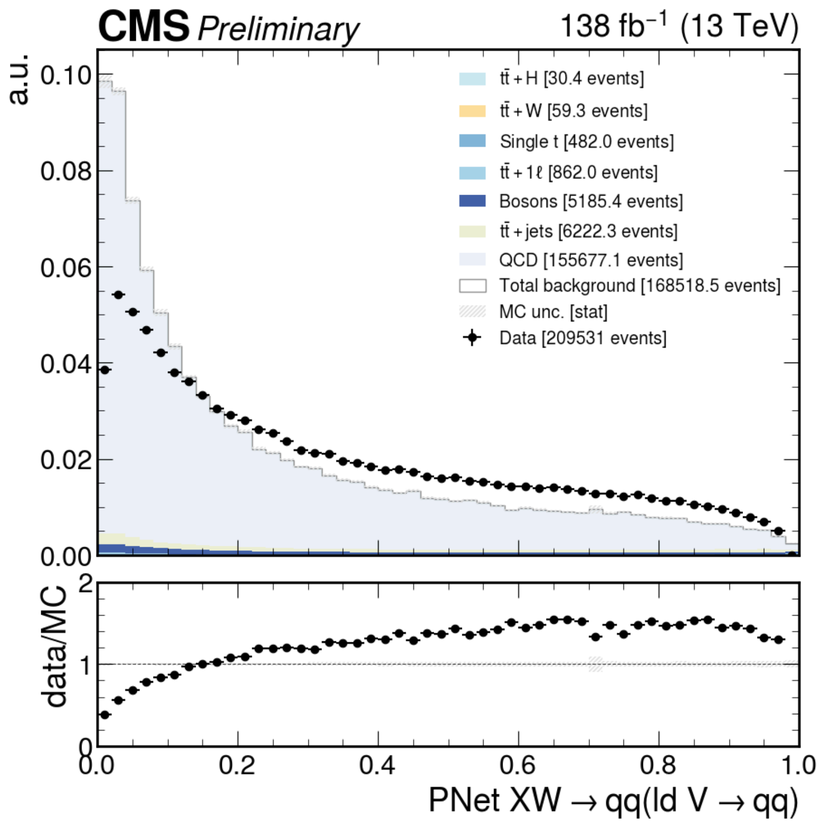
Yields scaled to lumi×σ, rounded for readability

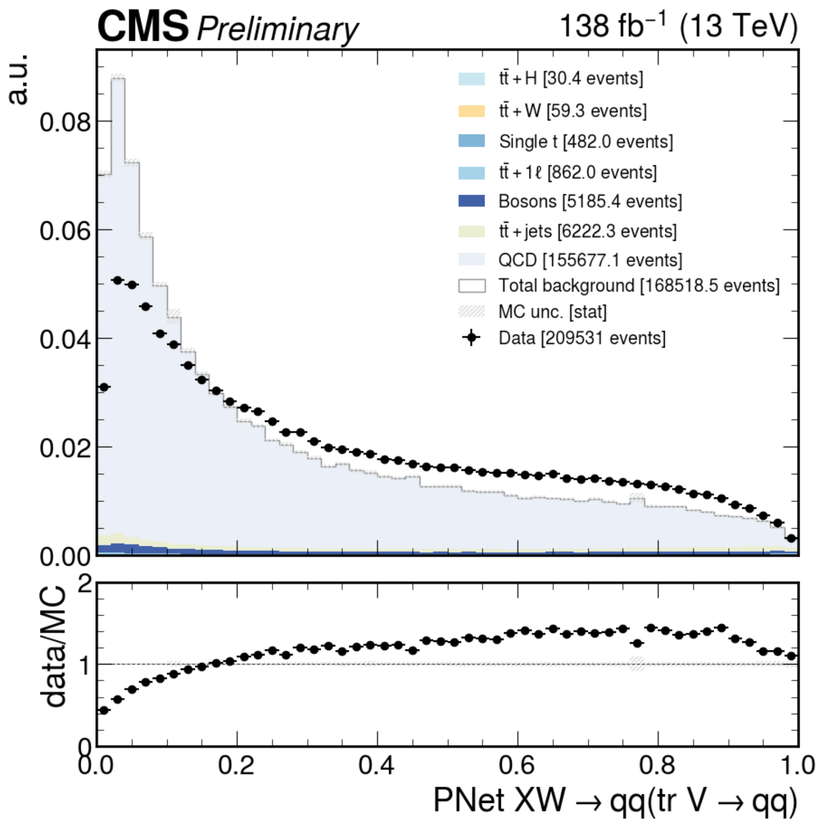
Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%

- Next: plot data/MC here and check for agreement
  - QCD is messy, so we expect some corrections will be needed
  - Safe to unblind since signal is so small

Data vs. MC: ParticleNet Scores







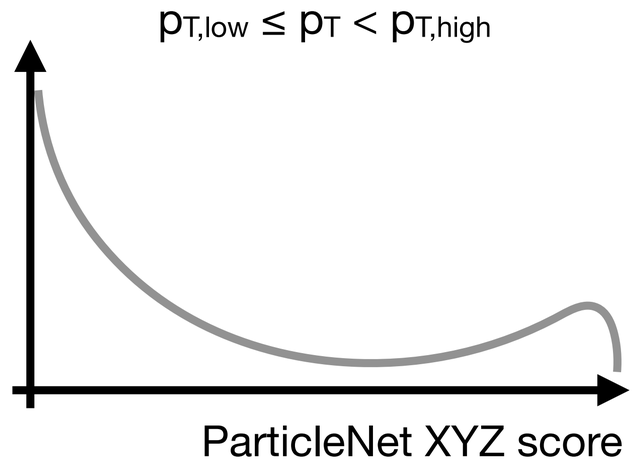
ParticleNet scores are not at all well modeled!

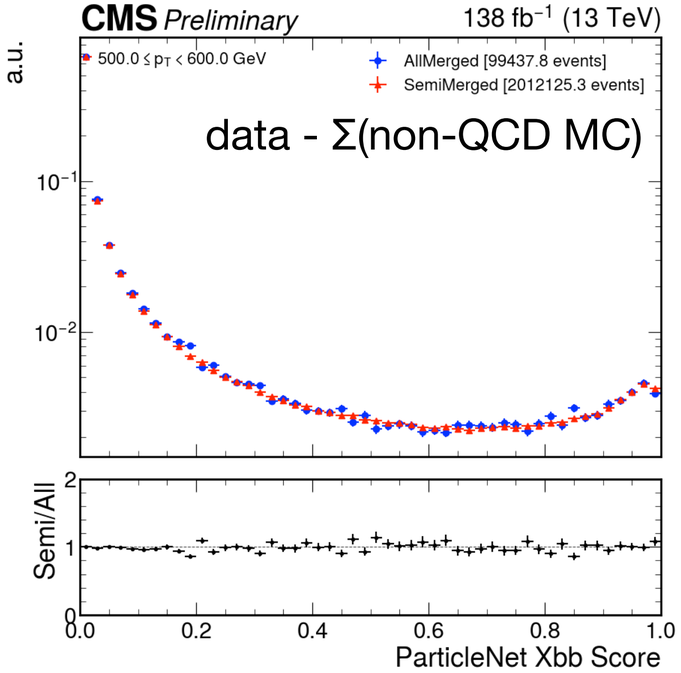
Needs some kind of correction: focus on QCD since it is the largest

UC San Diego

QCD Corrections

- ParticleNet scores are not well modeled by MC
- The ParticleNet XYZ* score for a given fat jet should be fundamentally described by a probability distribution function
- We can approximate this PDF by
  - Plotting the ParticleNet XYZ score **in data** for every fat jet in a histogram (really several: e.g. one per p_T bin)
  - Normalizing that histogram to unity
- The PDF should be the same for fat jets in the **3 fat jet channel** (main) vs. **2 fat jet channel** ✓
- **Goal:** replace MC ParticleNet scores in 3 fat jet channel with those sampled from the 2-fat jet PDF (**from data**)





*XYZ = Xbb, XWqq, etc.

UC San Diego

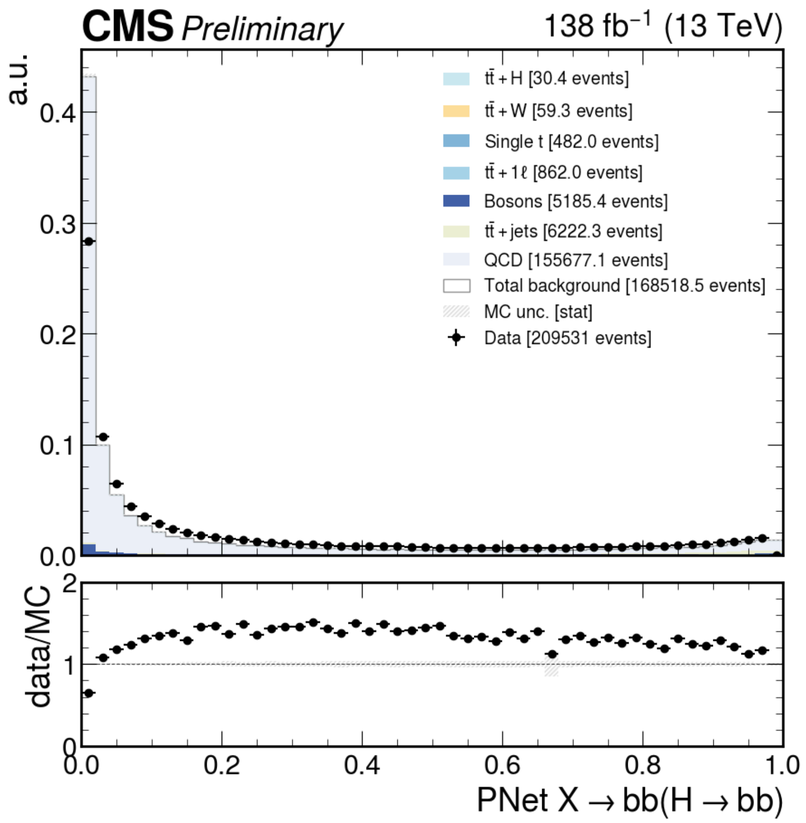
QCD Corrections

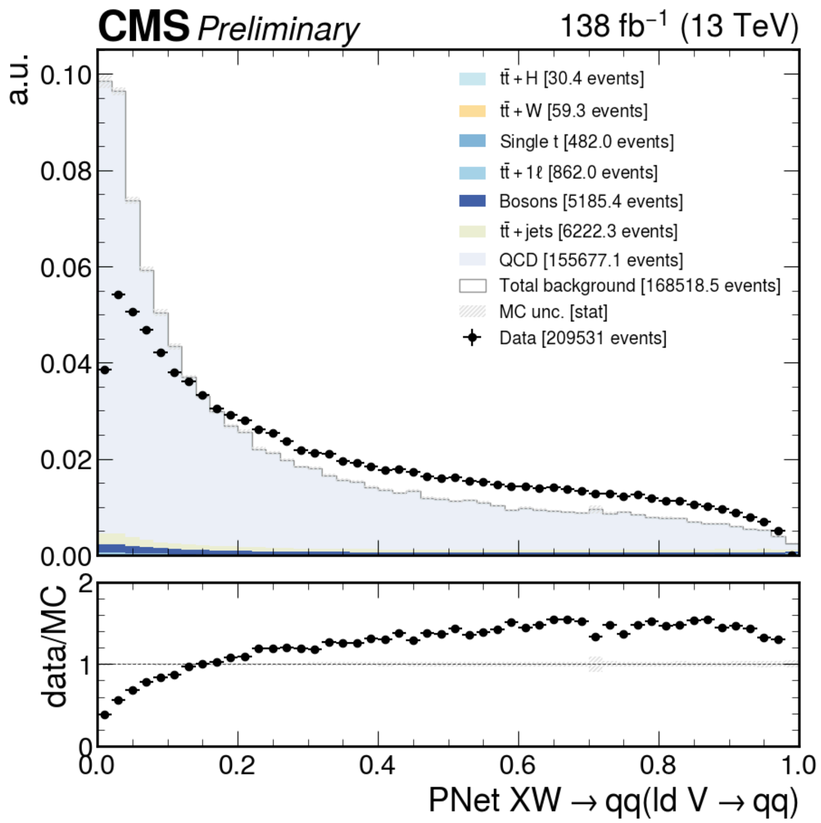
- All plots can be found here
- ParticleNet scores are replaced as follows (pseudocode):

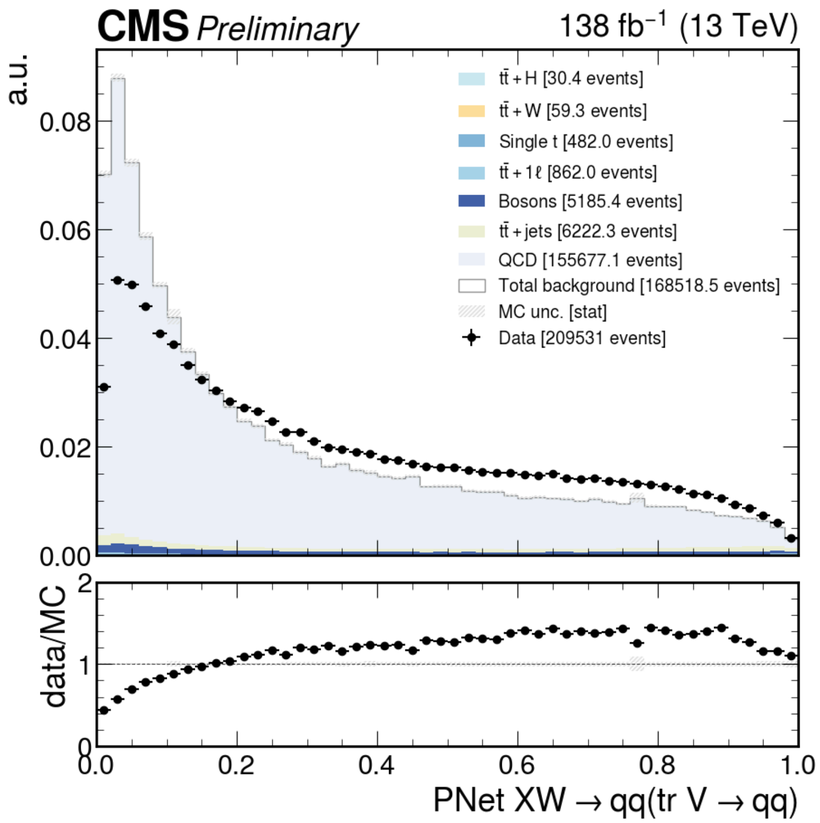
```
for fatjet in fatjets:
    // Sample PDFs for Xbb and XWqq scores
    pt = fatjet.pt()
    xbb = xbb_pdf2D.ProjectionY(pt).GetRandom()
    xwqq = xwqq_pdf3D.ProjectionZ(pt, xbb).GetRandom()
    // Replace Xbb and XWqq scores
    fatjet.xbb = xbb
    fatjet.xwqq = xwqq
```

- **The “projection” methods are pseudocode shorthand** for getting a slice of an N-dimensional histogram along only one axis in one bin of the others
- XWqq PDF is binned in p_T and Xbb score because $H \to b\bar{b}$ candidate is selected by max(Xbb) before the $V \to qq$ candidates are selected

QCD Corrections: Before



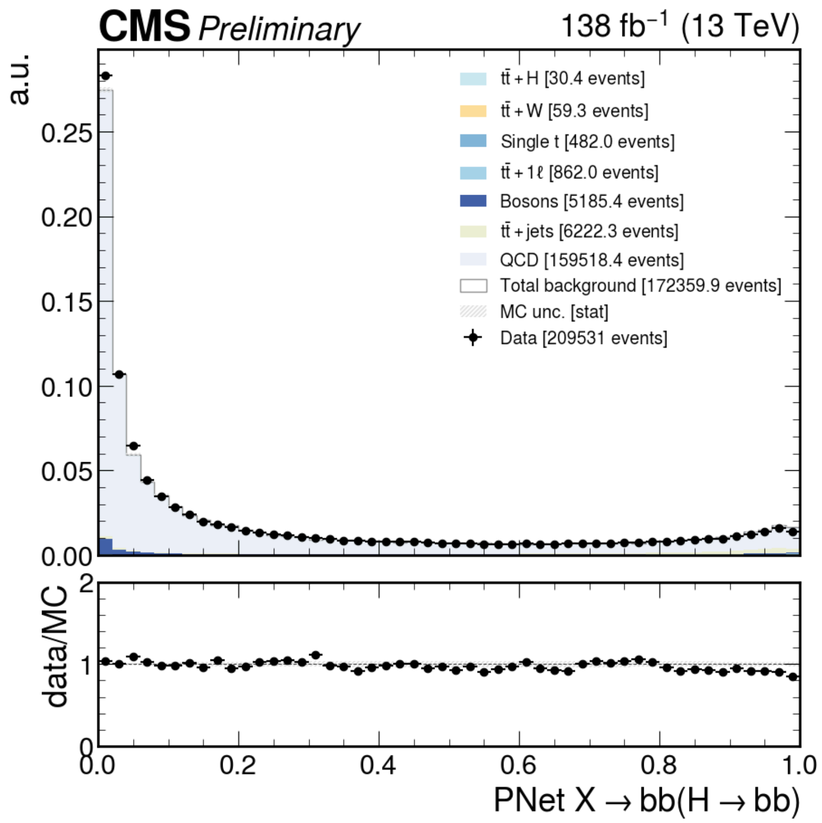


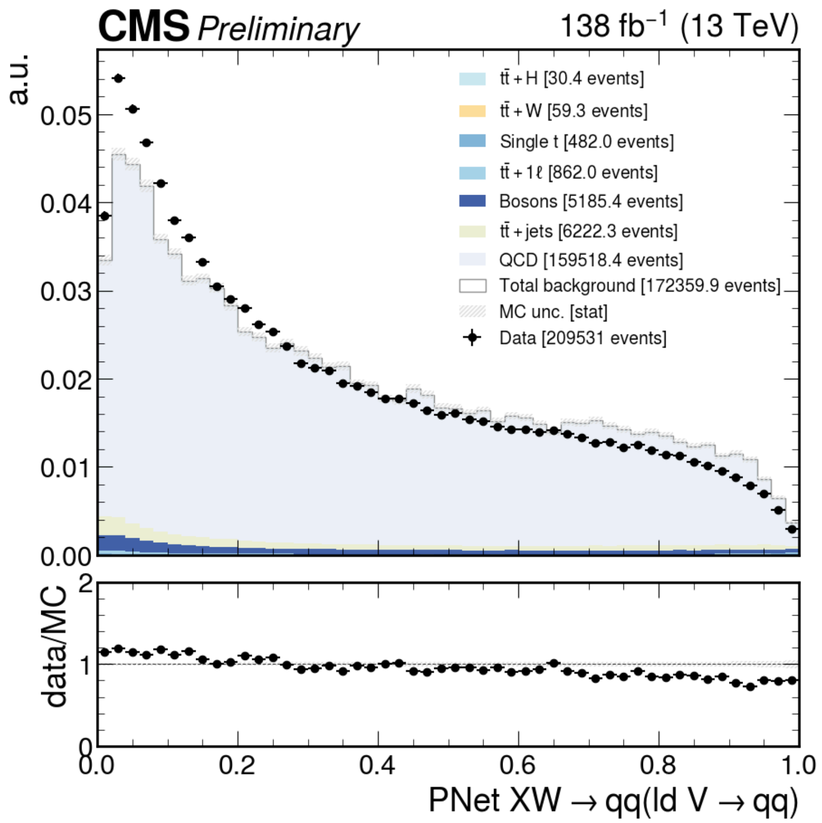


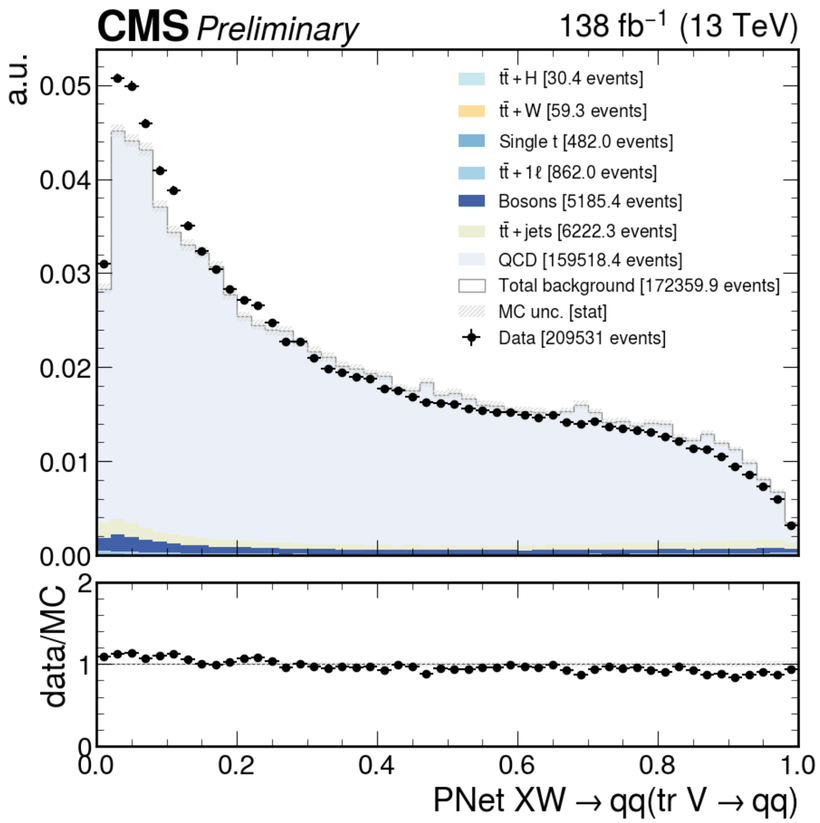
ParticleNet scores are not at all well modeled!

UC San Diego

QCD Corrections: After



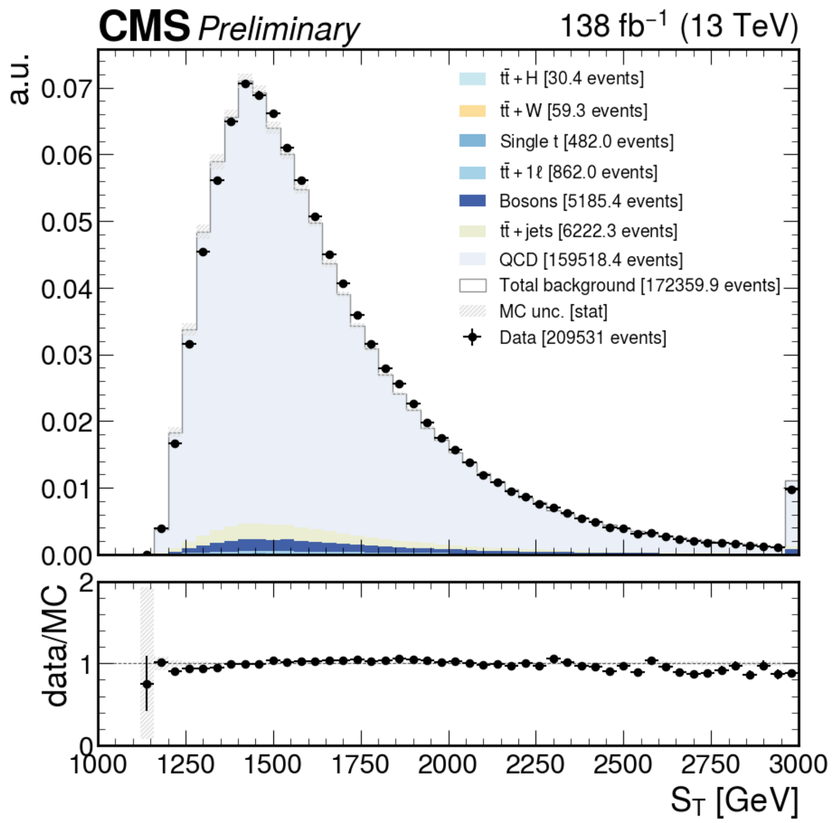


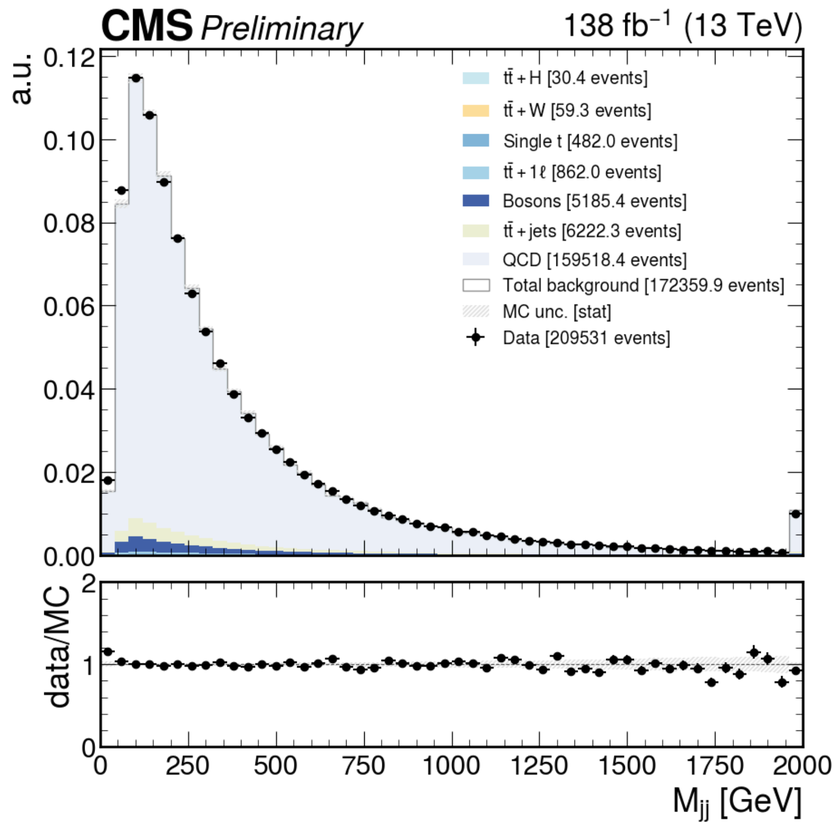


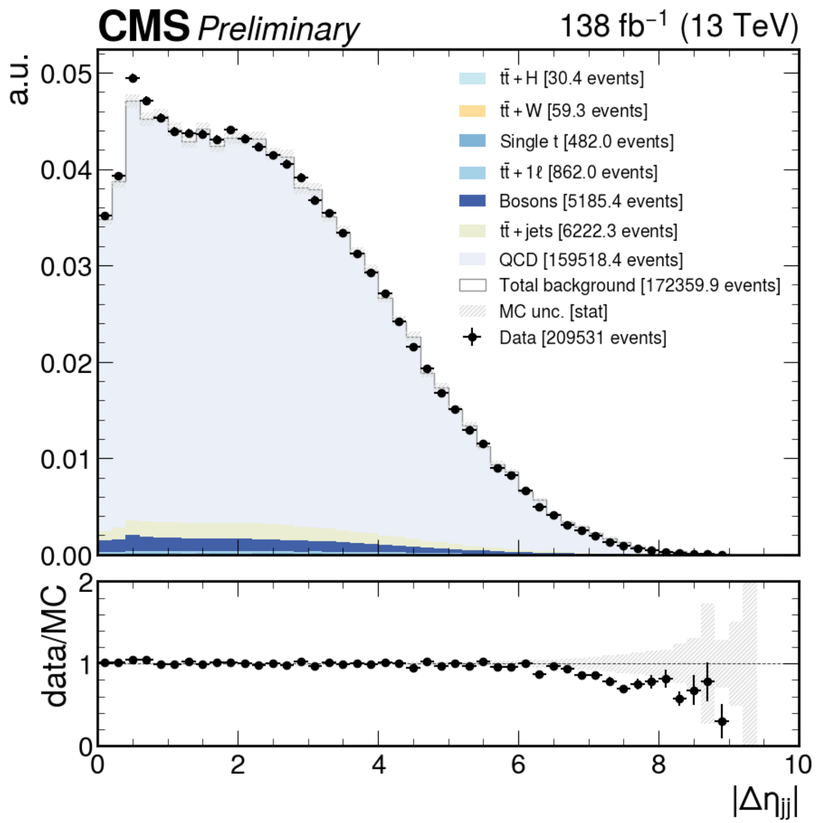
Not perfect, but shapes agree within 20% → ParticleNet scores are now usable!

UC San Diego

QCD Corrections: After







Other variable shapes look fine (more plots here)

Finally: rescale QCD integral to data - Σ(non-QCD MC)

UC San Diego

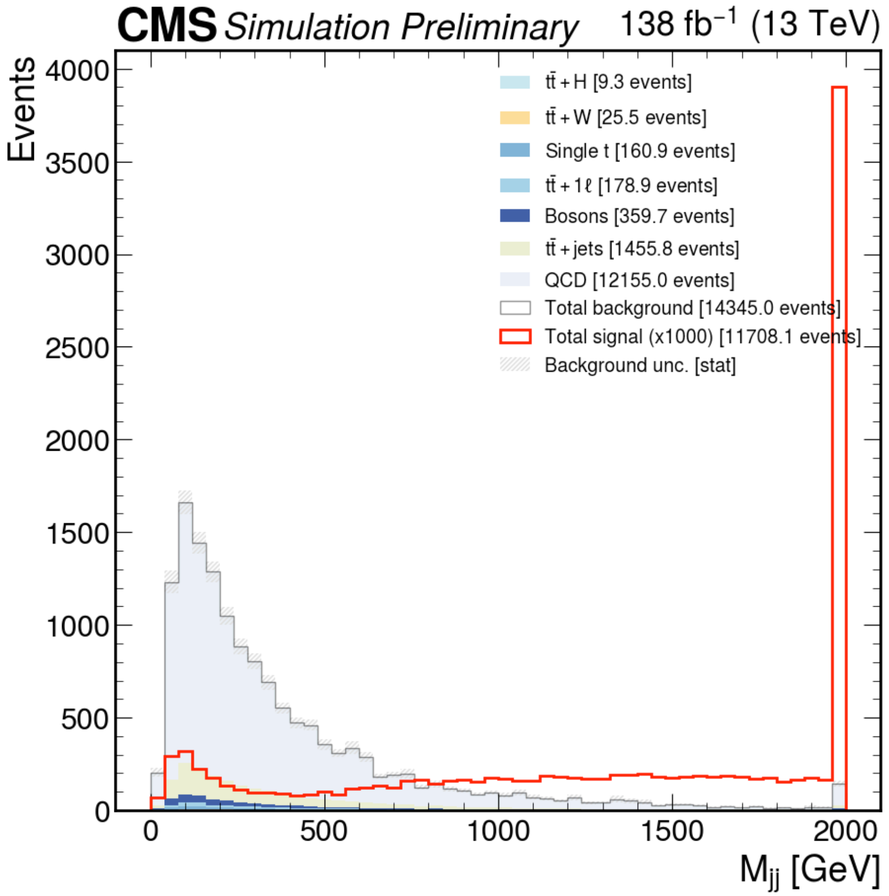
Preselection

Yields scaled to lumi×σ, rounded for readability

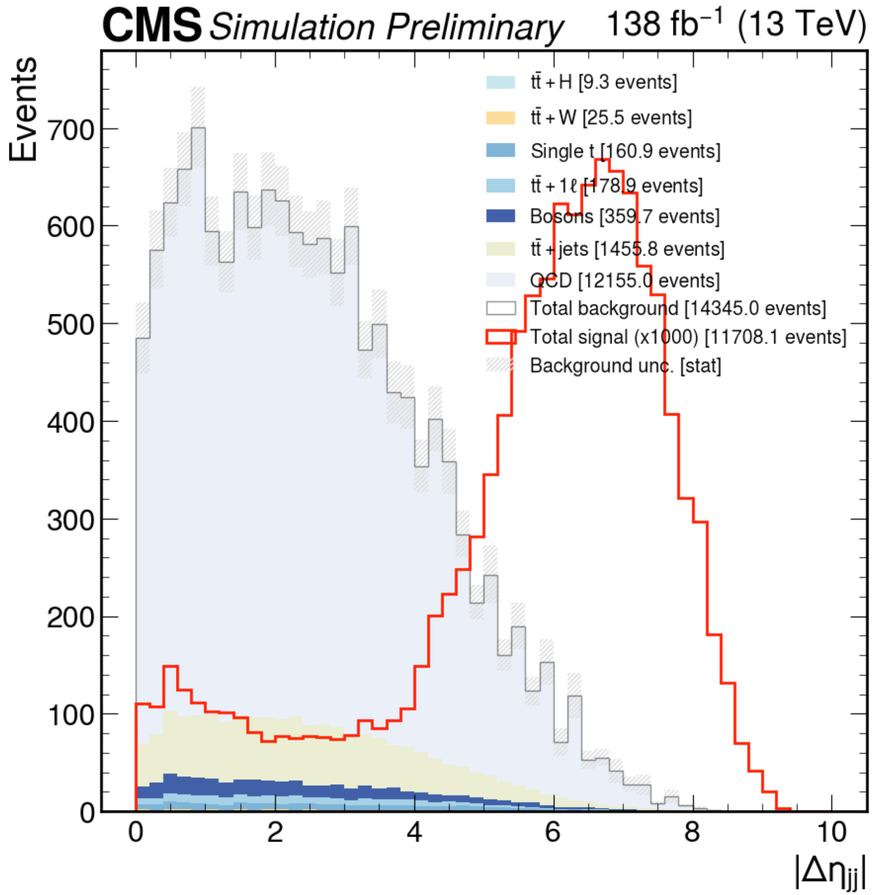
Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%
Preselection	12K	1.5K	179	25	9	161	360	14K	92%	12	34%

- Now that data/MC is better, we can trust MC shape more than before
- Make loose selection on ParticleNet scores to walk us slightly closer to a SR
  - Xbb($H \to b\bar{b}$) > 0.5 and XWqq(ld $V \to qq$) > 0.3 and XWqq(tr $V \to qq$) > 0.3
  - Recall: we rescale QCD integral to data - Σ(non-QCD MC)
- **Next:** plot important variables sig. vs. bkg. on next slides

VBS Variables (Preselection)



M_{jj} = Mass(ld VBS jet p4 + tr VBS jet p4)

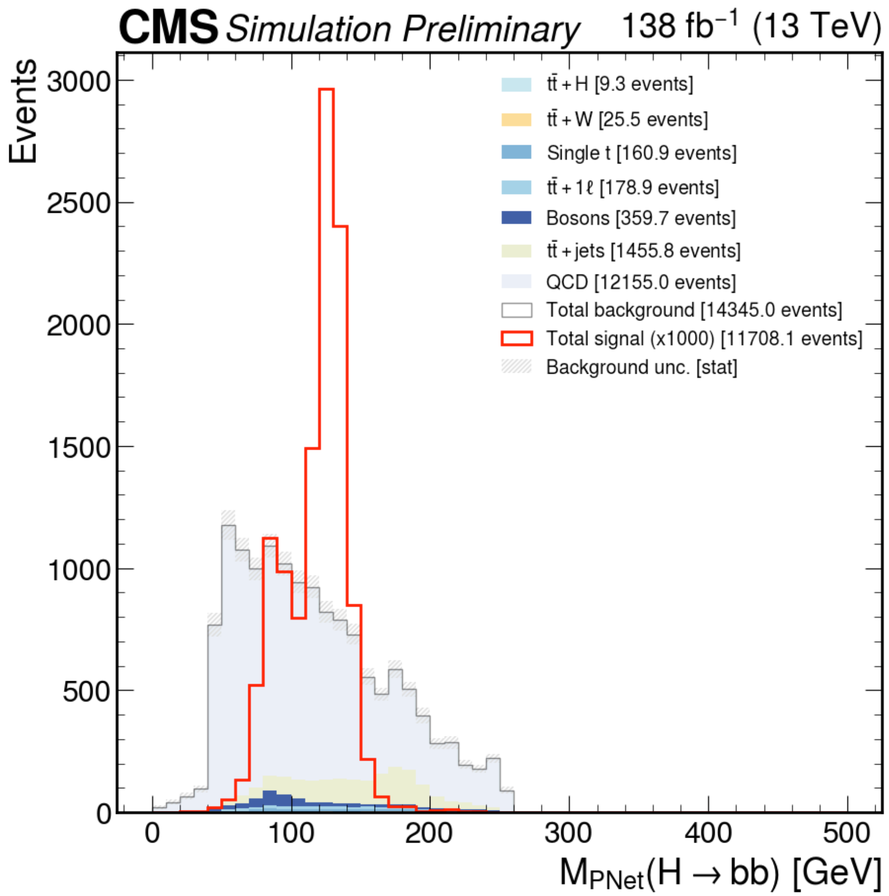


$|\Delta\eta_{jj}|$ = |ld VBS jet η - tr VBS jet η|

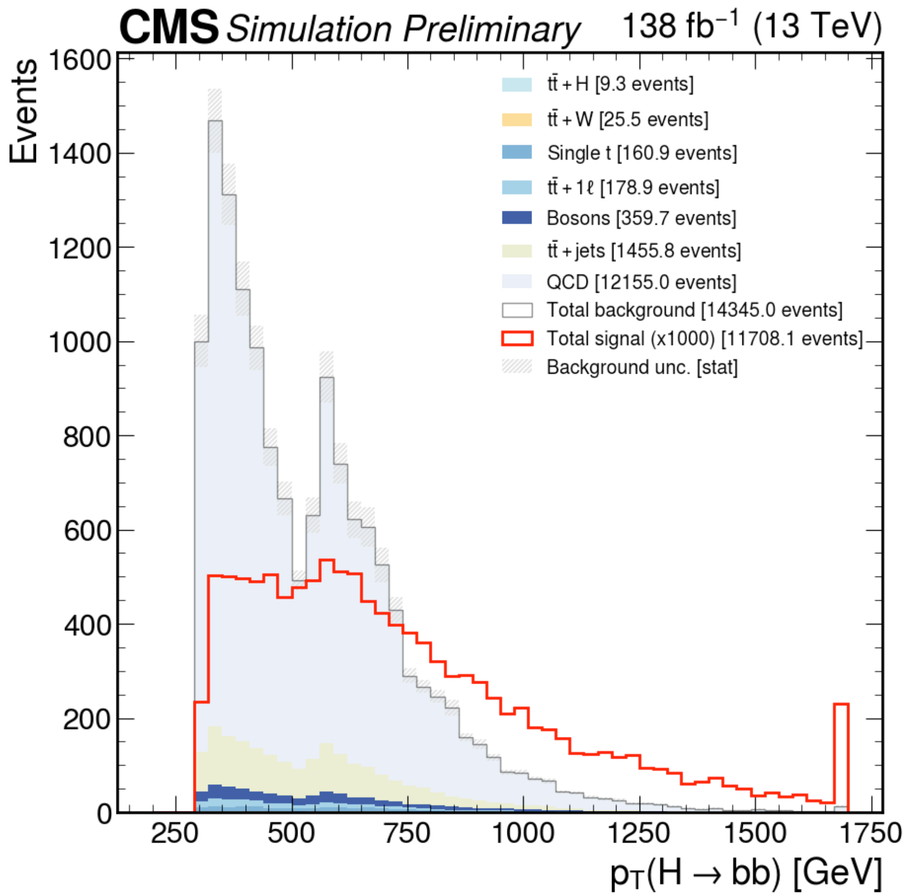
Characteristically large $\Delta\eta_{jj}$ and M_{jj} for signal (C_{2V} = 2)

UC San Diego

H→bb Variables (Preselection)



M_{PNet} = ParticleNet regressed mass

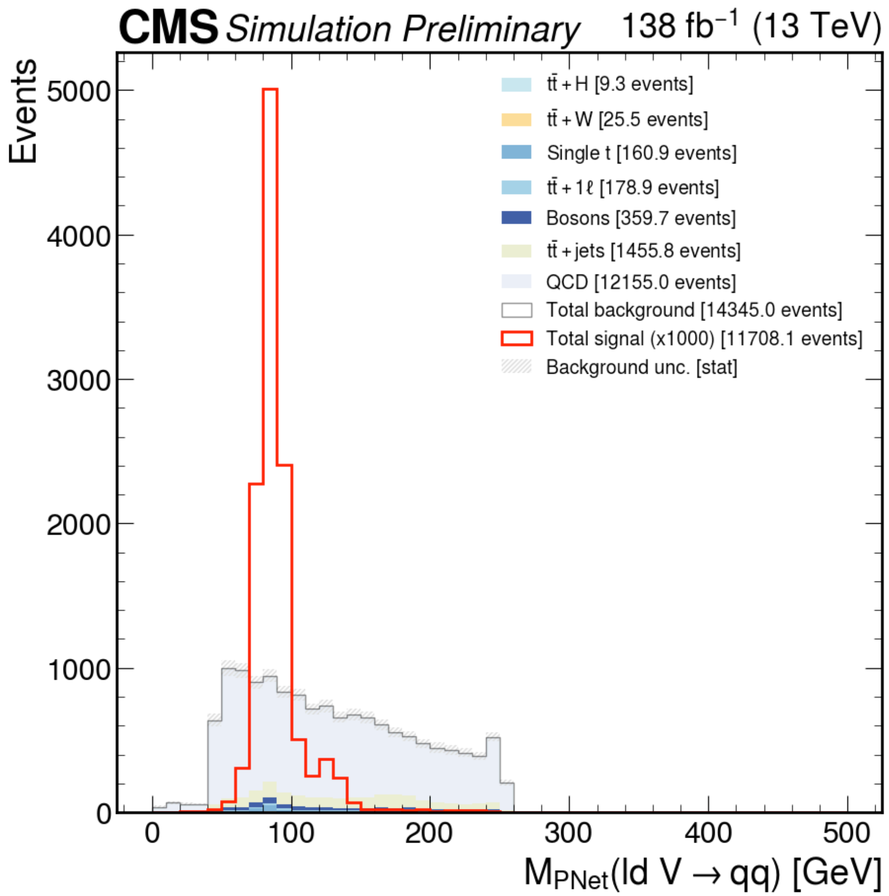


Spike at 500 GeV due to HLT threshold cut

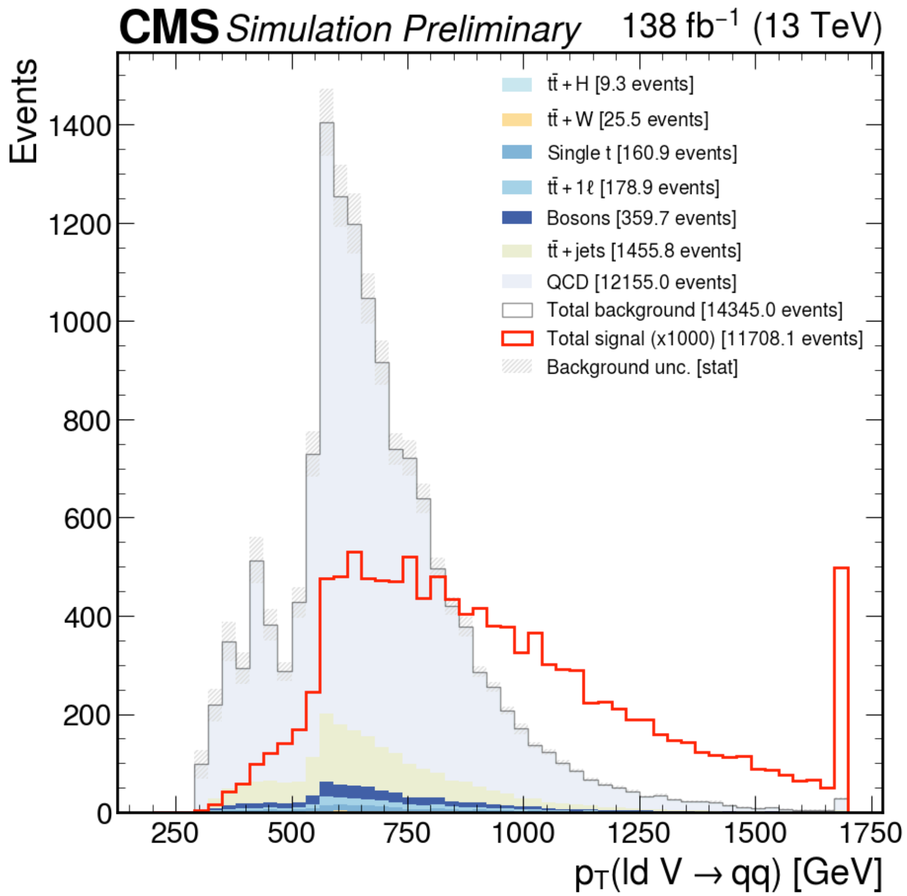
Higgs peak in regressed mass + large p_T for signal (C2V = 2)

UC San Diego

V→qq Variables (Preselection)



M_{PNet} = ParticleNet regressed mass

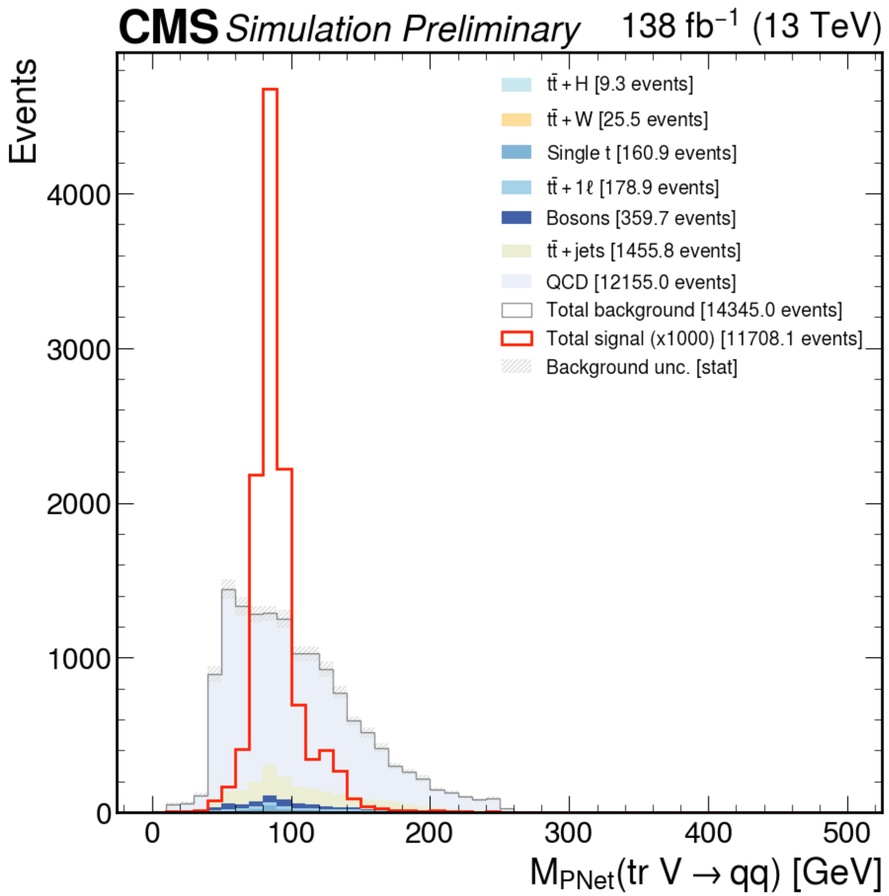


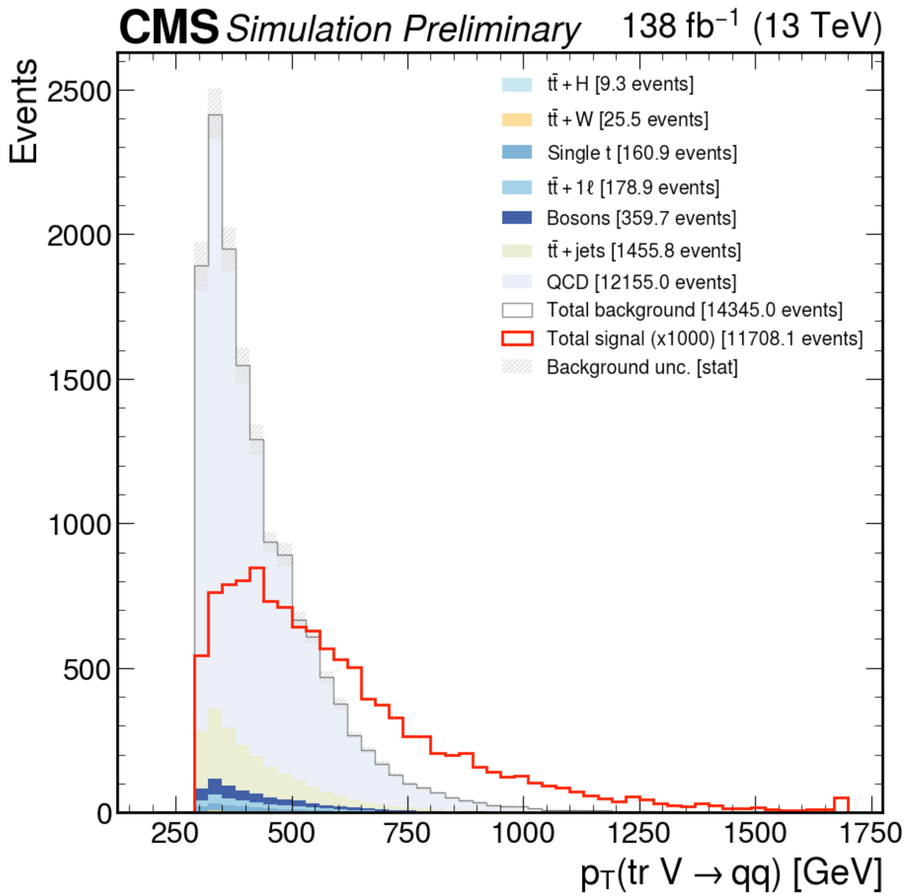
Spike at 500 GeV due to HLT threshold cut

W/Z peak in regressed mass + large p_T for signal (C2V = 2)

UC San Diego

V→qq Variables (Preselection)





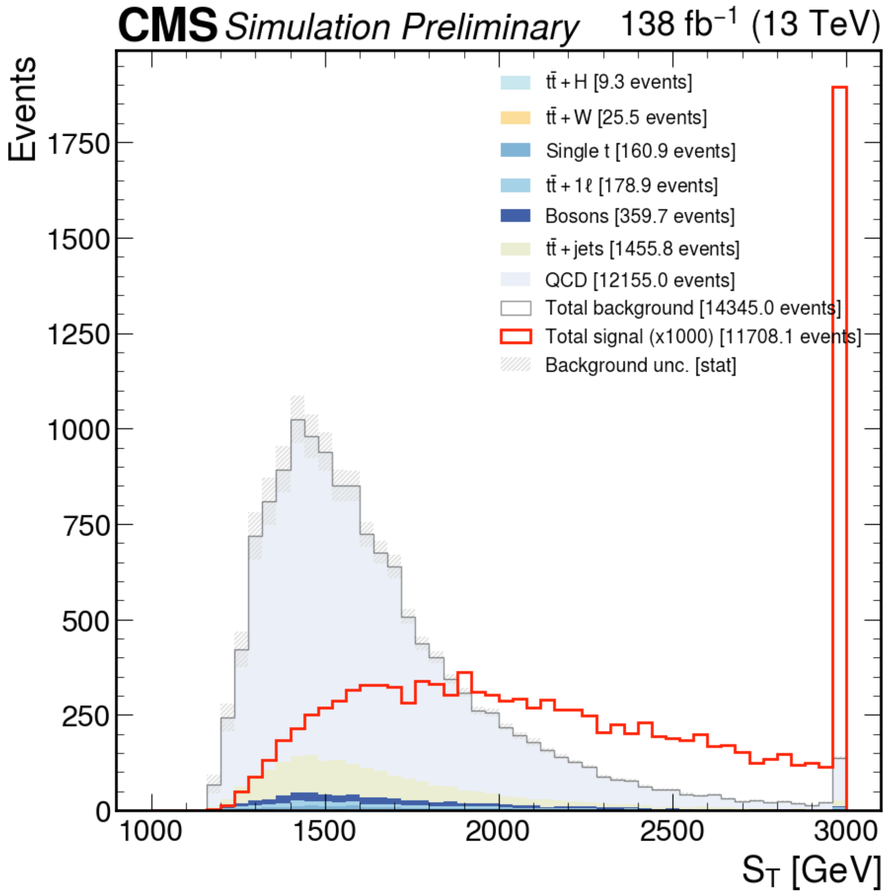
M_{PNet} = ParticleNet regressed mass

Trailing (tr) in p_T ⇒ no spike at 500 GeV from cut

W/Z peak in regressed mass + large p_T for signal (C2V = 2)

UC San Diego

Other Variables (Preselection)



$S_T = p_T(H \rightarrow bb) + p_T(\text{ld } V \rightarrow qq) + p_T(\text{tr } V \rightarrow qq)$

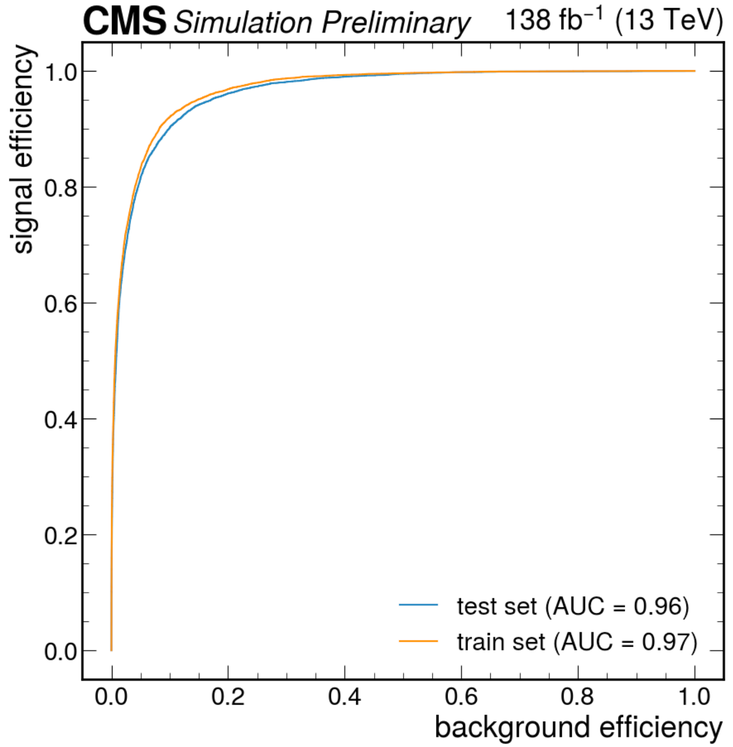
Expectedly large S_T for signal (C2V = 2)

BDT-based Signal Region

Yields scaled to lumi×σ, rounded for readability

Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%
Preselection	12K	1.5K	179	25	9	161	360	14K	92%	12	34%

- Train a simple BDT to get a “ceiling” for analysis sensitivity
- Use the following inputs:
  - H→bb fat jet p4 (p_T, η, ϕ), M_{PNet}
  - V→qq fat jet p4 (p_T, η, ϕ), M_{PNet}
  - VVH system p_T
- BDT hyperparameters tabulated in backup

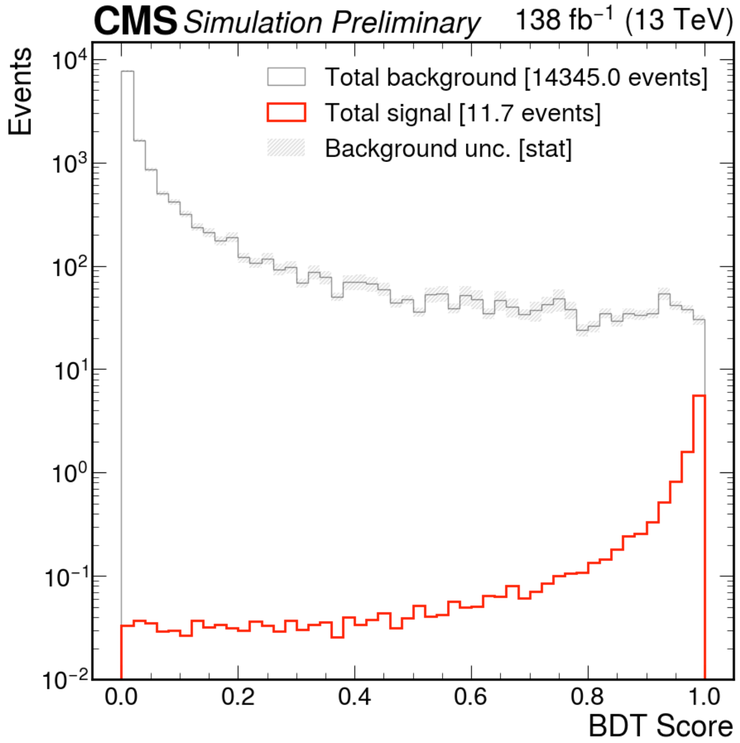


BDT-based Signal Region

Yields scaled to lumi×σ, rounded for readability

Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%
Preselection	12K	1.5K	179	25	9	161	360	14K	92%	12	34%

- Brute-force scan over the following cuts:
  - BDT score
  - ParticleNet scores (Xbb, XWqq)
  - M_{jj}
  - $|\Delta\eta_{jj}|$
- Optimize for $S/\sqrt{B}$ as a significance heuristic



UC San Diego

BDT-based Signal Region

NOT USED

Yields scaled to lumi×σ, rounded for readability

Cut	QCD	$t\bar{t}$+jets	$t\bar{t}$+1ℓ	$t\bar{t}$+W	$t\bar{t}$+H	Single top	Bosons	Total Bkg.	Eff.	VBSVVH (C_{2V} = 2)	Eff.
Skim	137,061K	748K	86K	2.6K	1.3K	53K	1,513K	139,464K	—	175	—
HLT + MET Filters	88,702K	575K	70K	2.2K	1.1K	41K	1,120K	90,512K	35%	168	4%
At least 3 fat jets	395K	9.8K	1.4K	110	46	874	13K	421K	100%	32	81%
Object selection	158K	6.2K	855	59	30	478	5.1K	171K	59%	18	44%
Preselection	12K	1.5K	179	25	9	161	360	14K	92%	12	34%
Signal Region	0.14	0.37	0.04	-0.02	0.00	0.08	0.18	0.81	100%	5	57%

- Settled on the following signal region:
  BDT > 0.9 and Xbb > 0.5 and XWqq > 0.82|0.66 (ld|tr) and $|\Delta\eta_{jj}| > 4$ and $M_{jj} > 600$ GeV
- Therefore the “ceiling” is high: **5 signal vs. 1 background for $C_{2V} = 2$**
- Next step: background extrapolation
  - We would like it to be data-driven, as we do not trust QCD
  - **Spoiler:** ABCD does not work with the BDT, so we pivot to a novel technique

Background Extrapolation

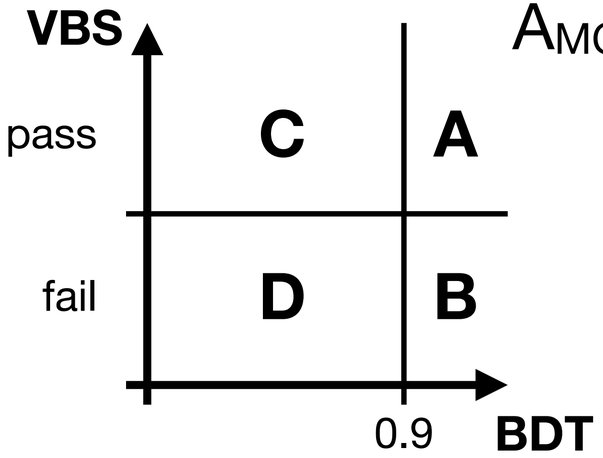
ABCD: BDT Signal Region

NOT USED

Xbb > 0.50 and XWqq > 0.82|0.66 (ld|tr)

Selection	Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
$\|\Delta\eta_{jj}\| > 4$ and $M_{jj} > 600$ GeV and BDT > 0.9	A	**0.81**	0.28	5.05	0.06	—	—
$\|\Delta\eta_{jj}\| \leq 4$ and $M_{jj} \leq 600$ GeV and BDT > 0.9	B	20.87	2.62	1.03	0.03	25	5.00
$\|\Delta\eta_{jj}\| > 4$ and $M_{jj} > 600$ GeV and BDT ≤ 0.8	C	175.92	20.87	1.15	0.03	172	13.11
$\|\Delta\eta_{jj}\| \leq 4$ and $M_{jj} \leq 600$ GeV and BDT ≤ 0.9	D	1069.99	44.43	0.32	0.02	1190	34.50

Very performant SR, but terrible MC closure



$$A_{MC} = B_{MC} \times C_{MC} / D_{MC} = \mathbf{3.43 \pm 0.61}$$

$$\epsilon_{syst} = \left| 1 - \frac{A_{MC}^{pred}}{A_{MC}} \right| = \left| 1 - \frac{\mathbf{3.43}}{\mathbf{0.81}} \right| = 325.0\%$$

$$\epsilon_{stat} = \sqrt{\frac{1}{B_{data}} + \frac{1}{C_{data}} + \frac{1}{D_{data}}} = 21.6\%$$

Final Result

Expected sig. 5.05 ± 0.06

Predicted bkg. 3.61 ± 0.78 ± 11.7

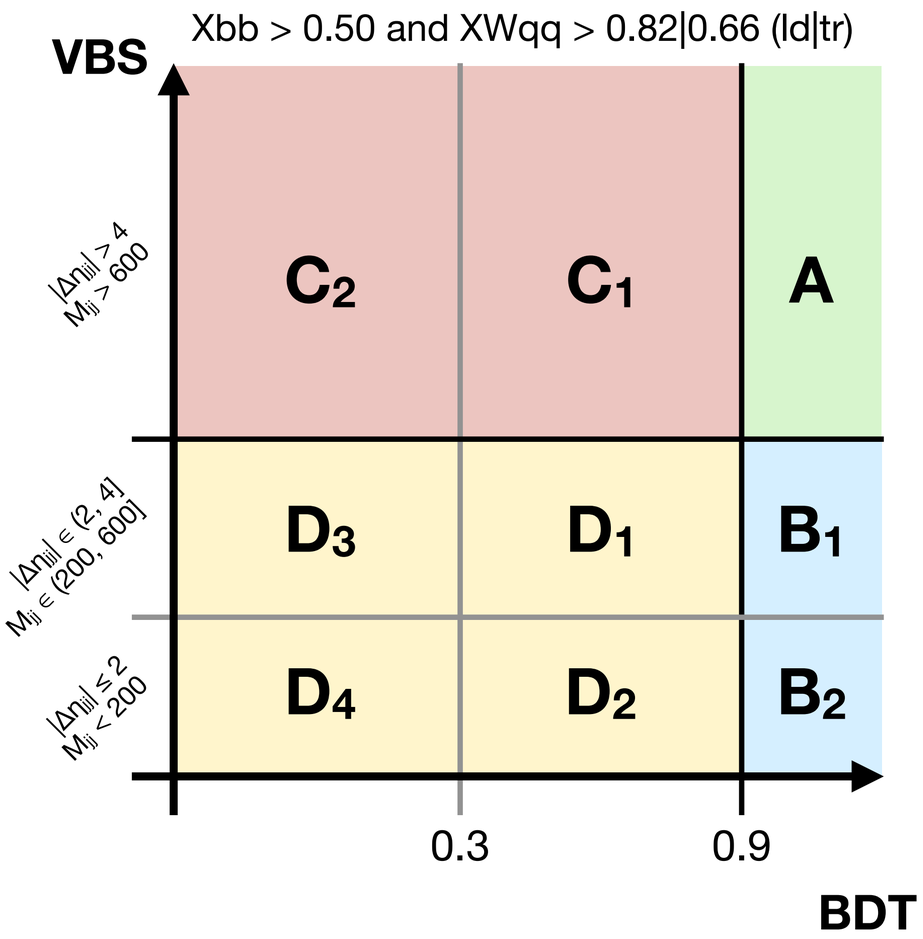
stat. *syst.*

Terrible closure ⇒ BDT & VBS correlated or poor MC composition modeling?

UC San Diego

BDT

NOT USED



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.81**	0.28	**5.05**	0.06	–	–
B	20.87	2.62	1.03	0.03	25	5.00
C	175.92	20.87	1.15	0.03	172	13.11
D	1070.0	44.43	0.32	0.02	1190	34.50

$$A^{pred} = B \times \frac{C}{D} = \begin{cases} 3.43 \pm 0.61 \text{ (MC)} \\ \mathbf{3.61 \pm 0.78} \text{ (Data)} \end{cases}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	12.42	1.85	0.59	0.02	**16**	4.00
B_2	8.44	1.86	0.44	0.02	9	3.00
D_1	93.31	12.75	0.15	0.01	87	9.33
D_2	43.38	4.16	0.09	0.01	55	7.42

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = \mathbf{14.24 \pm 5.34} \text{ (Data)} \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	93.31	12.75	0.15	0.01	**87**	9.33
D_2	43.38	4.16	0.09	0.01	55	7.42
D_3	604.79	34.56	0.04	0.01	649	25.48
D_4	328.51	24.49	0.03	0.00	399	19.97

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{89.46 \pm 13.3} \text{ (Data)} \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	19.72	7.16	1.01	0.03	**19**	4.36
D_1	93.31	12.75	0.15	0.01	87	9.33
C_2	156.20	19.61	0.13	0.01	153	12.37
D_3	604.79	34.56	0.04	0.01	649	25.48

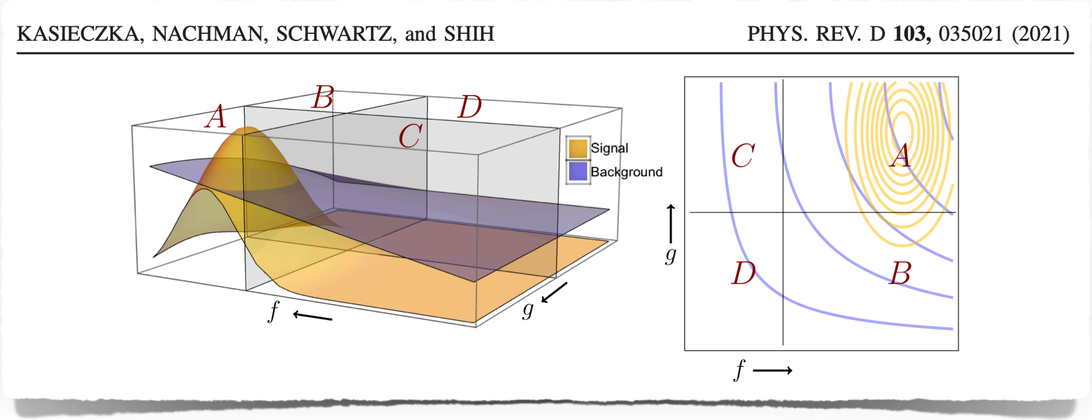
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{20.51 \pm 2.87} \text{ (Data)} \checkmark$$

ABCD works well in data (in sidebands), but predicted S/√B is not good

Studied for some time, then pivoted to “Automated ABCD” (next slides)

UC San Diego

Automated ABCD



- Need the following:
  - Performant signal region A
  - Uncorrelated “arms” f and g to perform bkg. extrapolation
- Previously using BDT as f, VBS cuts as g, but these become correlated closer to SR
- Enter: automated ABCD via ML (10.1103/PhysRevD.103.035021)
  - Train a deep neural network to serve as f and/or g
  - Add a decorrelation to the loss function that trains the network to be decorrelated from another variable
- In these slides: show **successful application of this technique for our analysis**

Automated ABCD

- Introduce Distance Correlation
  - DisCo for catchy titles, dCorr for math
  - dCorr(f, g) = 0 iff f and g are independent
  - dCorr(f, g) ∈ (0, 1] otherwise
- Claim high performance/easy to train vs. other decorrelation metrics
- Added to some typical loss (assume BCE)
- Hyperparameter λ controls relative size of DisCo term vs. BCE

PHYSICAL REVIEW D **103,** 035021 (2021)

Automating the ABCD method with machine learning

Gregor Kasieczka,[1,*] Benjamin Nachman,[2,†] Matthew D. Schwartz,[3,§] and David Shih[2,4,5,‡]
[1]*Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, D-22761 Hamburg, Germany*
[2]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[4]*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*
[5]*Berkeley Center for Theoretical Physics, University of California, Berkeley, California 94720, USA*

(Received 6 August 2020; accepted 3 February 2021; published 22 February 2021)

$$\mathcal{L}[f(X)] = \mathcal{L}_{\text{classifier}}[f(X), y] + \lambda \text{dCorr}^2_{y=0}[f(X), X_0], \quad (3.1)$$

$$\mathcal{L}[f, g] = \mathcal{L}_{\text{classifier}}[f(X), y] + \mathcal{L}_{\text{classifier}}[g(X), y] + \lambda \text{dCorr}^2_{y=0}[f(X), g(X)], \quad (3.2)$$

APPENDIX A: DISTANCE CORRELATION

For two random variables f and g, the distance covariance is defined as

$$\begin{aligned}\text{dCov}^2[f, g] = &\langle |f - f'| \times |g - g'| \rangle \\ &+ \langle |f - f'| \rangle \times \langle |g - g'| \rangle \\ &- 2\langle |f - f'| \times |g - g''| \rangle, \end{aligned} \quad (A1)$$

where (f, g), (f', g'), (f'', g'') are all independent and identically distributed from the same joint distribution. In practice, we evaluate $\text{dCov}^2[f, g]$ by averaging $|f_i - f_j| \times |g_i - g_j|$, $|f_i - f_j|$, and $|g_i - g_j|$ over all pairs of events i, j, and $|f_i - f_j| \times |g_i - g_k|$ over all triplets of events i, j, k.

The distance correlation is then defined analogously to the usual correlation:

$$\text{dCorr}^2[f, g] = \frac{\text{dCov}^2[f, g]}{\text{dCov}[f, f]\text{dCov}[g, g]}. \quad (A2)$$

Automated ABCD

- **Single DisCo:** train DNN to be decorrelated with some other variable X_0
  - i.e. DNN is f (we call it **ABCDNet**), X_0 is g
- **Double DisCo:** train two DNNs to be decorrelated with eachother
  - i.e. DNN1 is f, DNN2 is g
- Authors show three example applications:
  - 3D gaussians (reproduced in backup)
  - Top tagging
  - ATLAS SUSY analysis

PHYSICAL REVIEW D **103,** 035021 (2021)

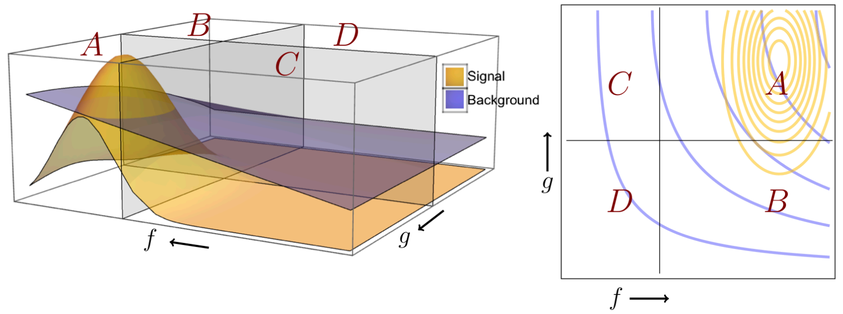
Automating the ABCD method with machine learning

Gregor Kasieczka,[1,*] Benjamin Nachman,[2,†] Matthew D. Schwartz,[3,§] and David Shih[2,4,5,‡]
[1]*Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, D-22761 Hamburg, Germany*
[2]*Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*
[4]*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*
[5]*Berkeley Center for Theoretical Physics, University of California, Berkeley, California 94720, USA*

(Received 6 August 2020; accepted 3 February 2021; published 22 February 2021)

$$\mathcal{L}[f(X)] = \mathcal{L}_{\text{classifier}}[f(X), y] + \lambda \text{dCorr}^2_{y=0}[f(X), X_0], \quad (3.1)$$

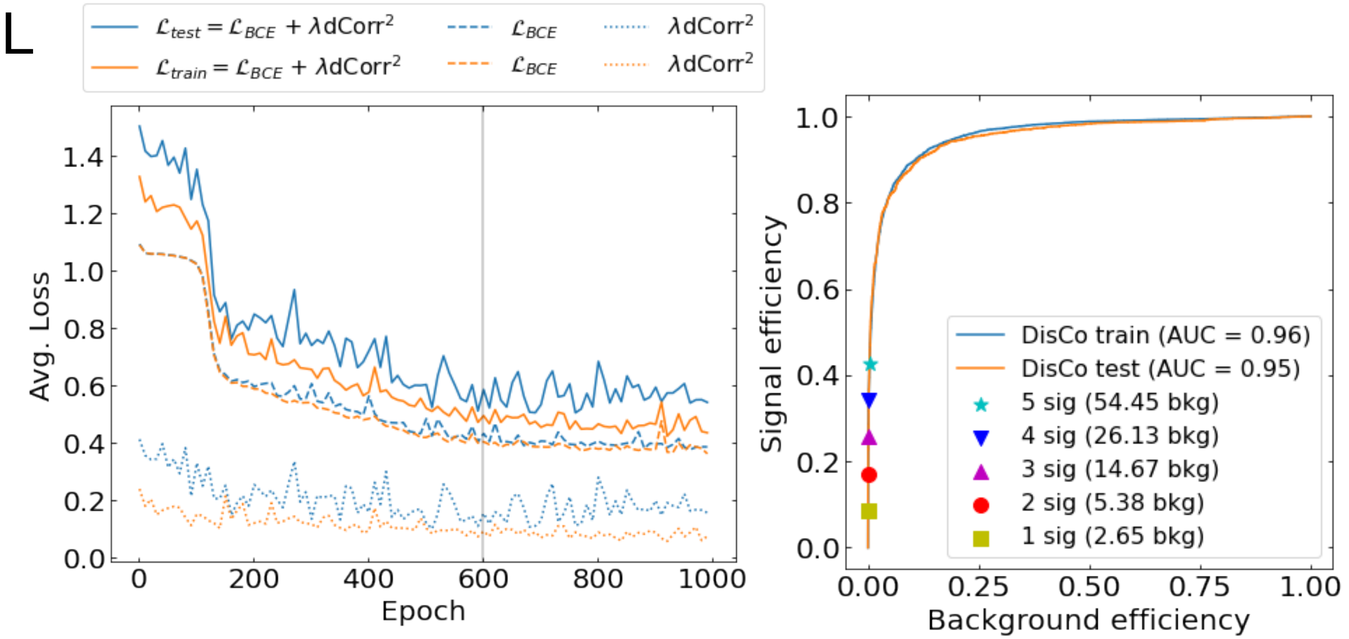
$$\mathcal{L}[f, g] = \mathcal{L}_{\text{classifier}}[f(X), y] + \mathcal{L}_{\text{classifier}}[g(X), y] + \lambda \text{dCorr}^2_{y=0}[f(X), g(X)], \quad (3.2)$$



ABCDNet: λ = 30 DisCo (Leaky ReLU)

$$\mathscr{L} = \mathscr{L}_{BCE}(f(\vec{x}), y) + 30 \times \text{dCorr}_{y=0}(f(\vec{x}), |\Delta\eta_{jj}|)$$

- Borrow ABCDNet architecture from PRL SUSY example
  - 3 hidden layers (64 nodes each)
- Input features:
  - H→bb fat jet p4 (i.e. p_T, η, ϕ), M_{PNet}
  - V→qq fat jet p4 (i.e. p_T, η, ϕ), M_{PNet}
  - M_{jj}
- p_T → ln(p_T), others → (x - min)/(max - min)



Tried a number of λ values:
λ > 30 destabilizes training (fails to converge)
λ < 30 misses decorrelation (poor ABCD closure)

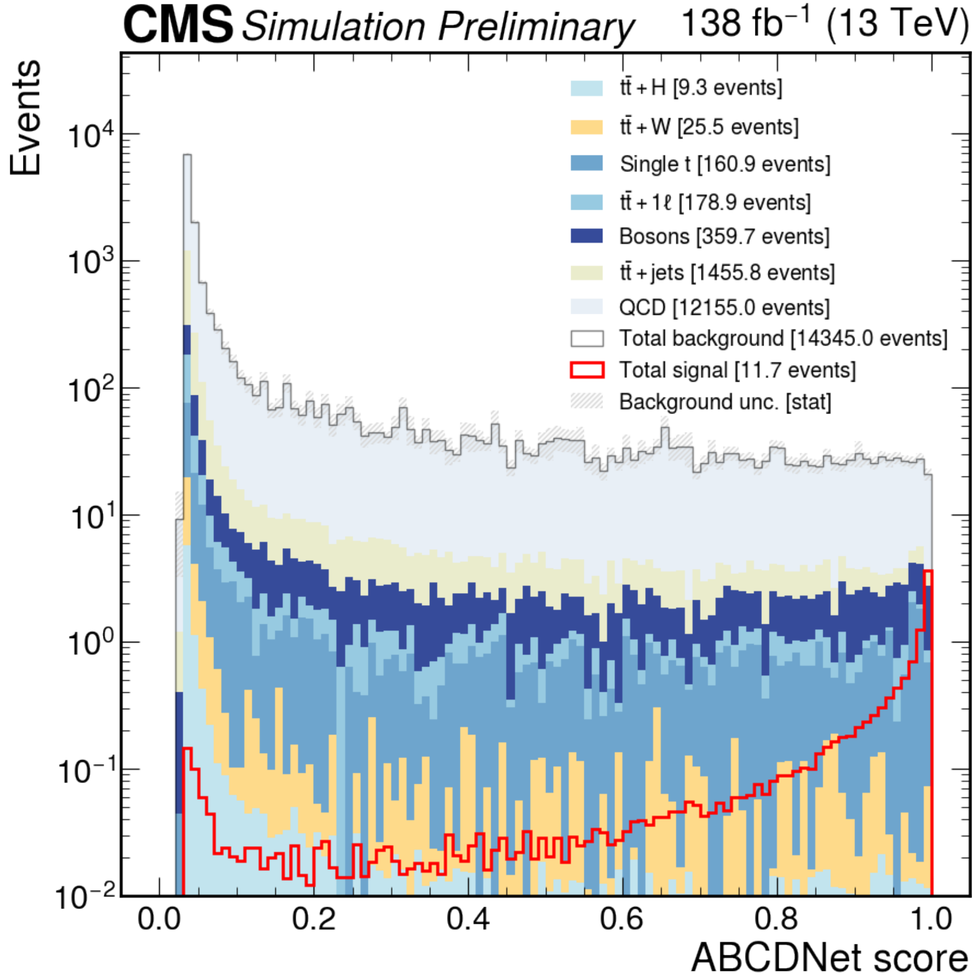
Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)

$$\mathscr{L} = \mathscr{L}_{BCE}(f(\vec{x}), y) + 30 \times \text{dCorr}_{y=0}(f(\vec{x}), |\Delta\eta_{jj}|)$$

- Perform brute-force scan over cuts on:
  - ABCDNet score
  - ParticleNet scores (Xbb, XWqq)
  - $|\Delta\eta_{jj}|$
- First, optimize for S/√B
- Finally, select slightly looser SR so there are sufficient stats in Region B of ABCD



Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

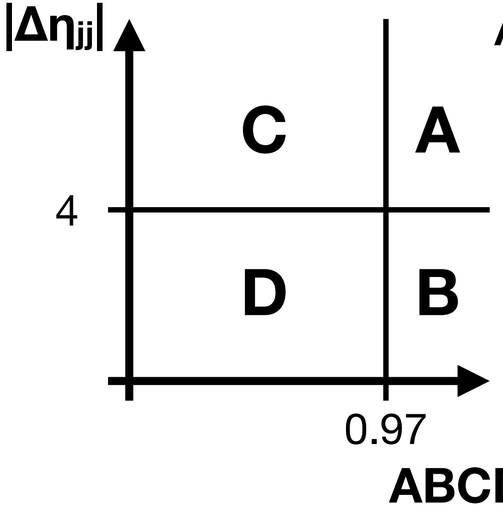
ABCDNet: λ = 30 DisCo (Leaky ReLU)

Xbb > 0.60 and XWqq > 0.75|0.70 (ld|tr)

Selection	Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
$\|\Delta\eta_{jj}\| > 4$ and ABCDNet > 0.97	A	1.82	0.57	3.48	0.05	—	—
$\|\Delta\eta_{jj}\| \leq 4$ and ABCDNet > 0.97	B	5.70	1.25	0.47	0.02	5	2.24
$\|\Delta\eta_{jj}\| > 4$ and ABCDNet ≤ 0.97	C	292.52	25.89	3.05	0.05	281	16.76
$\|\Delta\eta_{jj}\| \leq 4$ and ABCDNet ≤ 0.97	D	1011.83	41.07	0.65	0.02	1200	34.64

Fairly performant SR, and closure is within 1σ

(but Region D data/MC is not great)



$$A_{MC} = B_{MC} \times C_{MC} / D_{MC} = 1.65 \pm 0.40$$

$$\epsilon_{syst} = \left| 1 - \frac{A_{MC}^{pred}}{A_{MC}} \right| = \left| 1 - \frac{1.65}{1.82} \right| = 9.4\%$$

$$\epsilon_{stat} = \sqrt{\frac{1}{B_{data}} + \frac{1}{C_{data}} + \frac{1}{D_{data}}} = 45.2\%$$

Final Result

Expected sig. 3.48 ± 0.05

Predicted bkg. $1.17 \pm 0.53 \pm 0.11$

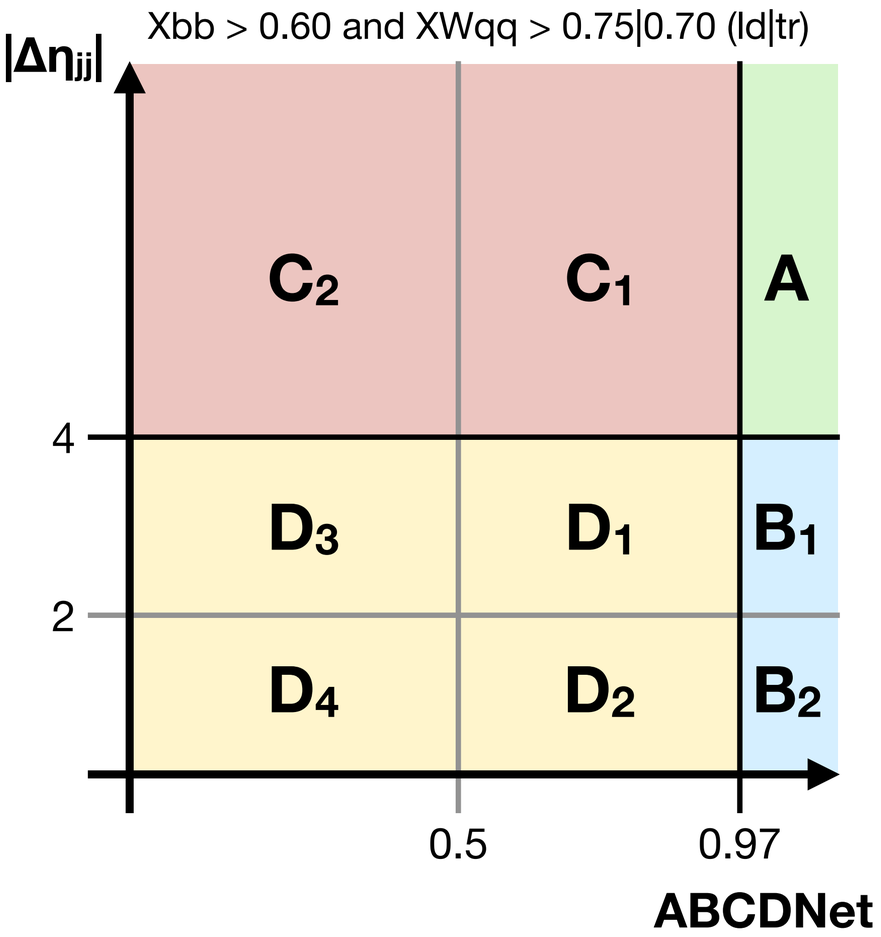
stat. *syst.*

Predicted significance (S/√B) and syst. error improved over previous results

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.82**	0.57	**3.48**	0.05	–	–
B	5.70	1.25	0.47	0.02	5	2.24
C	292.52	25.89	3.05	0.05	281	16.76
D	1011.83	41.07	0.65	0.02	1200	34.64

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.65 \pm 0.40 \text{ (MC)} \\ \mathbf{1.17 \pm 0.53} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	3.52	1.07	0.19	0.01	4	2.00
B_2	2.18	0.65	0.01	0.01	1	1.00
D_1	49.49	4.81	0.21	0.01	53	7.28
D_2	65.69	9.77	0.28	0.01	49	7.00

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 1.08 \pm 1.10 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	49.49	4.81	0.21	0.01	**53**	7.28
D_2	65.69	9.77	0.28	0.01	49	7.00
D_3	424.46	28.31	0.07	0.01	522	22.85
D_4	472.19	27.68	0.09	0.01	576	24.00

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{44.41 \pm 6.89} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	27.79	3.47	2.61	0.04	**29**	5.39
D_1	49.49	4.81	0.21	0.01	53	7.28
C_2	264.73	25.66	0.45	0.02	252	15.87
D_3	424.46	28.31	0.07	0.01	522	22.85

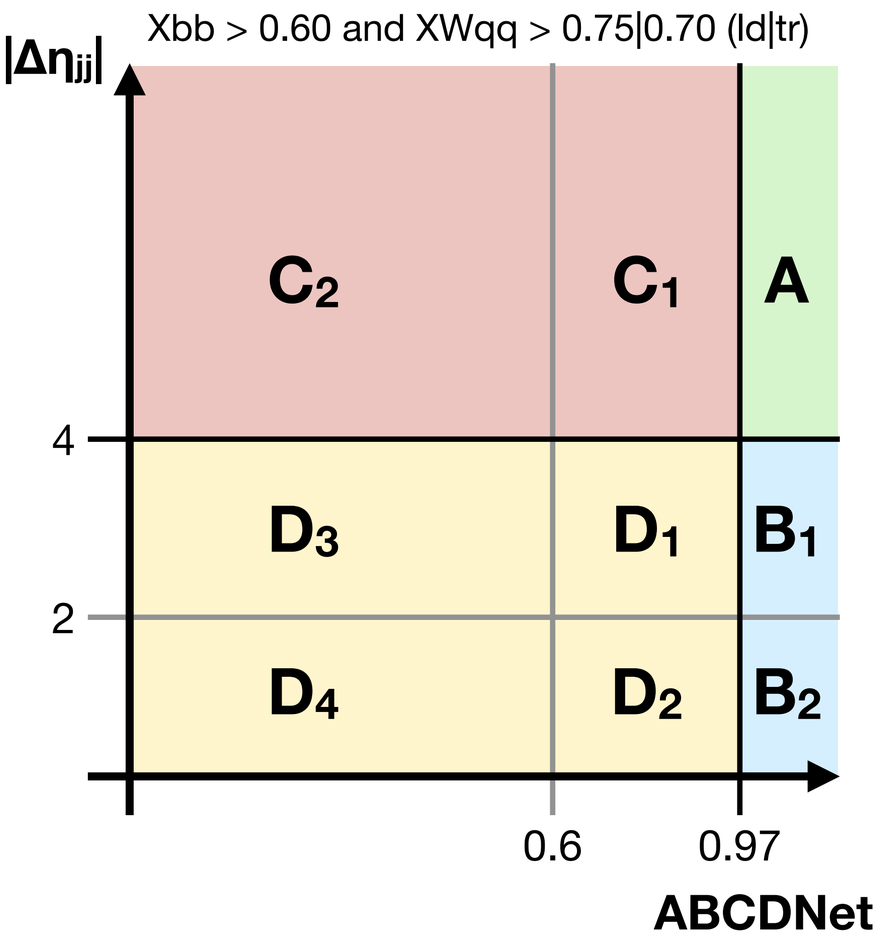
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{25.59 \pm 4.03} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.82**	0.57	**3.48**	0.05	–	–
B	5.70	1.25	0.47	0.02	5	2.24
C	292.52	25.89	3.05	0.05	281	16.76
D	1011.83	41.07	0.65	0.02	1200	34.64

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.65 \pm 0.40 \text{ (MC)} \\ \mathbf{1.17 \pm 0.53} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	3.52	1.07	0.19	0.01	4	2.00
B_2	2.18	0.65	0.01	0.01	1	1.00
D_1	35.57	3.89	0.20	0.01	38	6.16
D_2	56.41	9.58	0.26	0.01	38	6.16

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 1.00 \pm 1.03 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	35.57	3.89	0.20	0.01	**38**	6.16
D_2	56.41	9.58	0.26	0.01	38	6.16
D_3	438.38	28.45	0.08	0.01	537	23.17
D_4	481.47	27.75	0.11	0.01	587	24.23

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{34.76 \pm 6.01} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	19.76	2.88	2.47	0.04	**21**	4.58
D_1	35.57	3.89	0.20	0.01	38	6.16
C_2	272.76	25.73	0.58	0.02	260	16.12
D_3	438.38	28.45	0.08	0.01	537	23.17

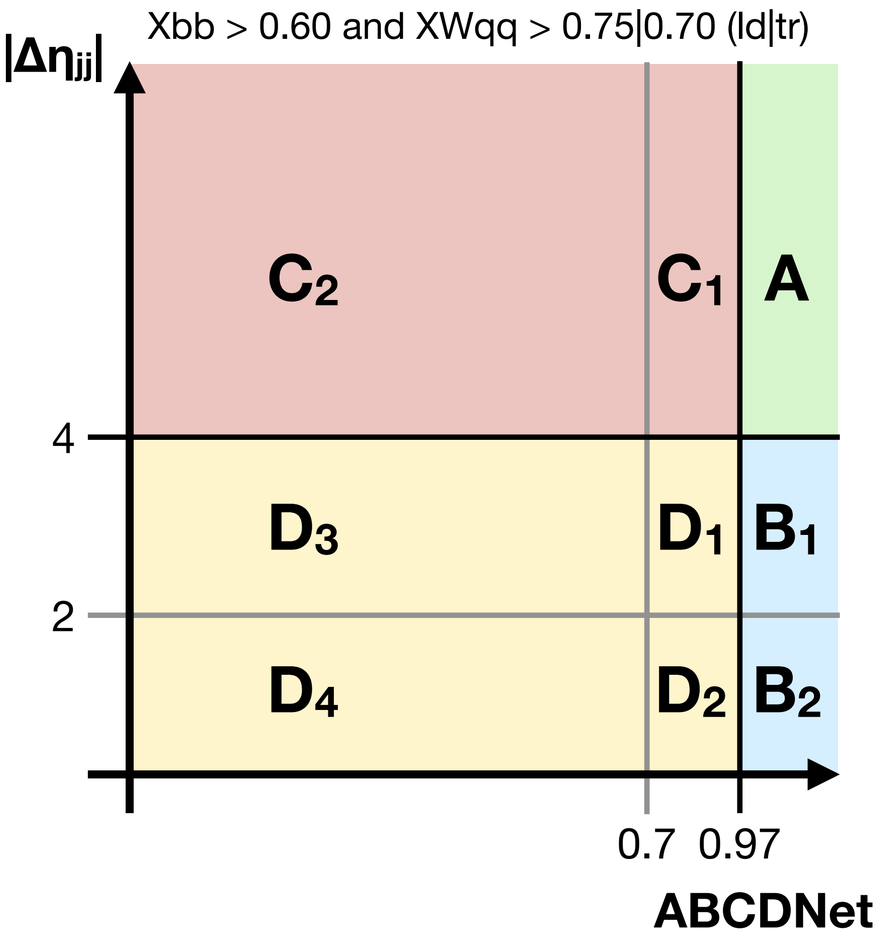
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{18.40 \pm 3.29} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.82**	0.57	**3.48**	0.05	–	–
B	5.70	1.25	0.47	0.02	5	2.24
C	292.52	25.89	3.05	0.05	281	16.76
D	1011.83	41.07	0.65	0.02	1200	34.64

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.65 \pm 0.40 \text{ (MC)} \\ \mathbf{1.17 \pm 0.53} \text{ (Data)} \end{bmatrix}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	3.52	1.07	0.19	0.01	4	2.00
B_2	2.18	0.65	0.11	0.01	1	1.00
D_1	28.88	3.52	0.18	0.01	29	5.39
D_2	38.67	7.42	0.23	0.01	27	5.20

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 1.07 \pm 1.11 \text{ (Data) ?}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	28.88	3.52	0.18	0.01	**29**	5.39
D_2	38.67	7.42	0.23	0.01	27	5.20
D_3	445.08	28.50	0.10	0.01	546	23.37
D_4	499.21	28.40	0.14	0.01	598	24.45

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{24.65 \pm 4.96} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	14.65	2.37	2.28	0.04	**14**	3.74
D_1	28.88	3.52	0.18	0.01	29	5.39
C_2	277.86	25.78	0.78	0.02	267	16.34
D_3	445.08	28.50	0.10	0.01	546	23.37

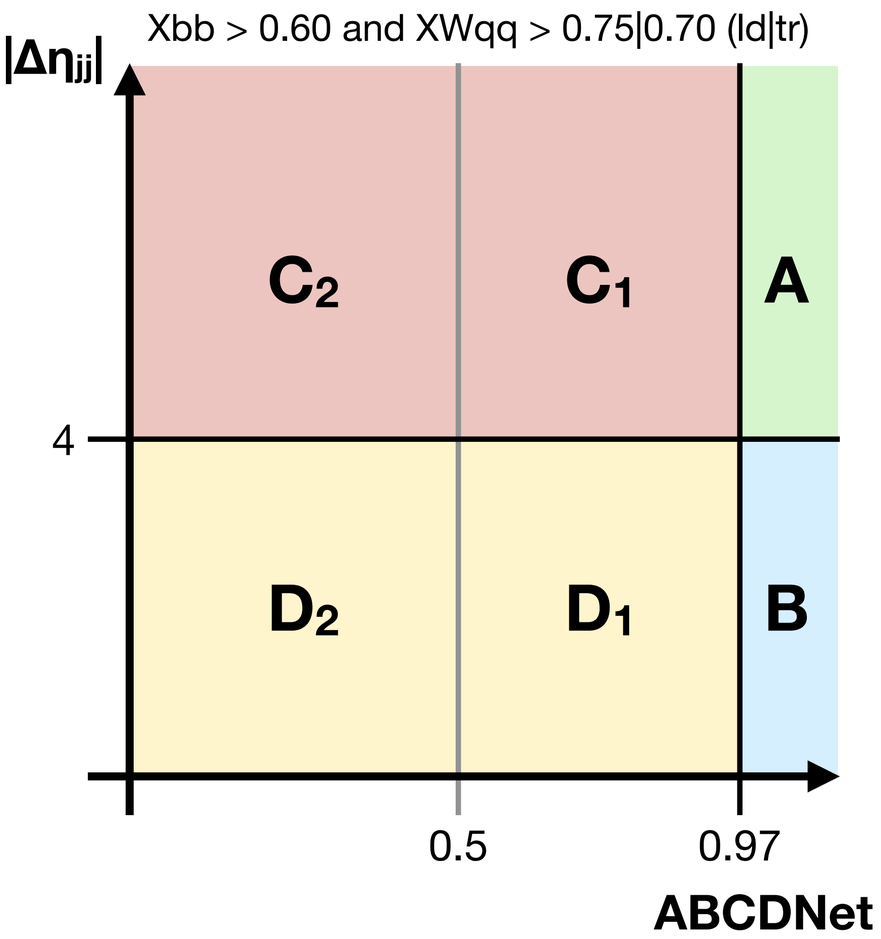
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{14.18 \pm 2.84} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.82**	0.57	**3.48**	0.05	–	–
B	5.70	1.25	0.47	0.02	5	2.24
C	292.52	25.89	3.05	0.05	281	16.76
D	1011.8	41.07	0.65	0.02	1200	34.64

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.65 \pm 0.40 \text{ (MC)} \\ \mathbf{1.17 \pm 0.53} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.82**	0.57	**3.48**	0.05	–	–
B	5.70	1.25	0.47	0.02	5	2.24
C_1	27.79	3.47	2.61	0.04	29	5.39
D_1	115.2	10.89	0.49	0.02	102	10.10

$$A^{pred} = B \times \frac{C_1}{D_1} = \begin{bmatrix} 1.37 \pm 0.37 \text{ (MC)} \\ \mathbf{1.42 \pm 0.70} \text{ (Data)} \end{bmatrix}$$

✓

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	27.79	3.47	2.61	0.04	**29**	5.39
D_1	115.18	10.89	0.49	0.02	102	10.10
C_2	264.73	25.66	0.45	0.02	252	15.87
D_2	896.65	39.60	0.16	0.01	1098	33.14

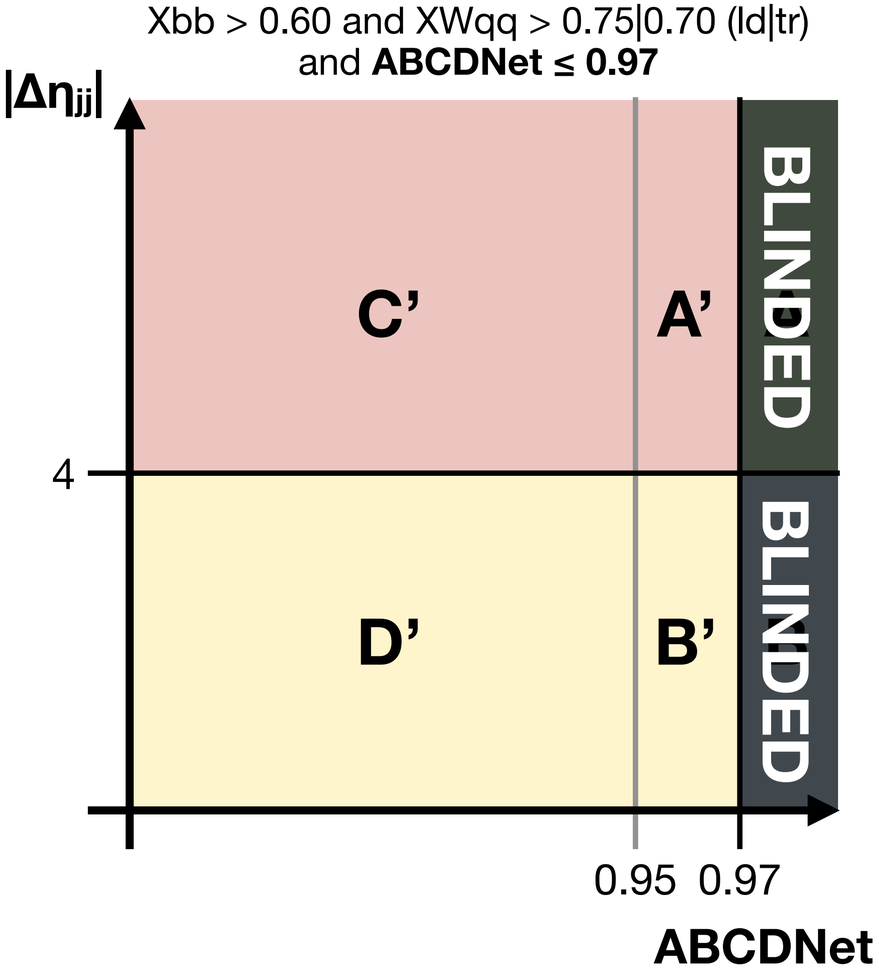
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{23.41 \pm 2.84} \text{ (Data)} \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A'	**1.66**	0.88	0.53	0.02	1	1.0
B'	5.74	1.35	0.09	0.01	4	2.00
C'	290.86	25.88	2.52	0.04	280	16.73
D'	1006.1	41.04	0.56	0.02	1196	34.58

$$A^{pred} = B \times \frac{C}{D} = \begin{cases} \mathbf{1.66 \pm 0.42} \text{ (MC)} \\ \mathbf{0.94 \pm 0.47} \text{ (Data)} \end{cases}$$

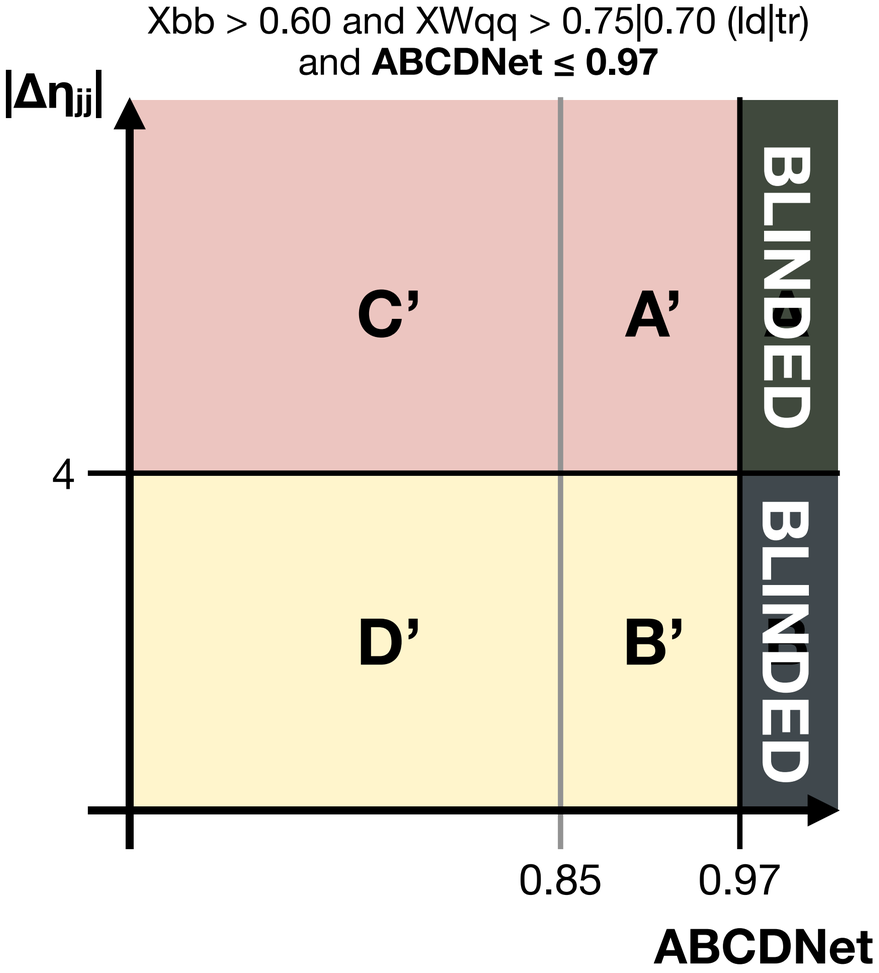
- Blind ABCDNet SR cut
- Perform ABCD (here: A'B'C'D') to predict number of background in a "near-SR" part of the C+D sideband
- Closer to "real-life" closure test in **data**

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A'	**6.96**	1.66	1.73	0.04	7	2.65
B'	27.66	3.38	0.30	0.02	23	4.80
C'	285.55	25.84	1.32	0.03	274	16.55
D'	984.2	40.93	0.35	0.02	1177	34.31

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} \mathbf{8.03 \pm 1.27}\ \text{(MC)} \\ \mathbf{5.35 \pm 1.17}\ \text{(Data)} \end{bmatrix}$$

- Blind ABCDNet SR cut
- Perform ABCD (here: A'B'C'D') to predict number of background in a "near-SR" part of the C+D sideband
- Closer to "real-life" closure test in **data**
  - Loosen "near-SR" for more statistics

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

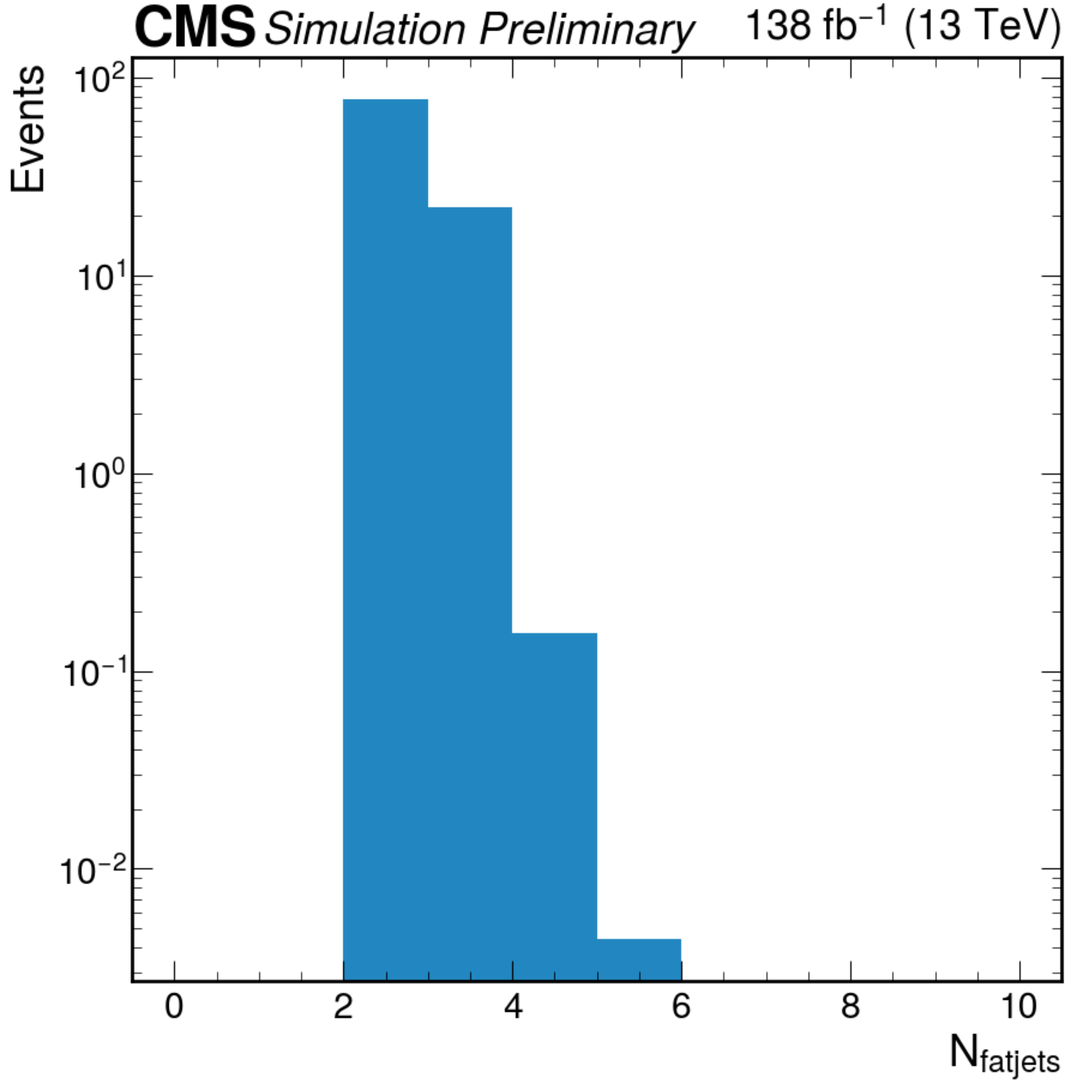
Summary

- We apply a correction to the QCD MC so it better resembles data
- The λ = 30 Single DisCo model seems to work well
  - MC closure is acceptable + good data closure in sidebands
  - **This method seems usable**
- Next steps:
  - More supporting material/studies to support the bkg. extrapolation method?
  - Scale factors (ParticleNet, Pileup ID, …) and signal systematics
  - Final limit

Backup

$N_{fatjets}$

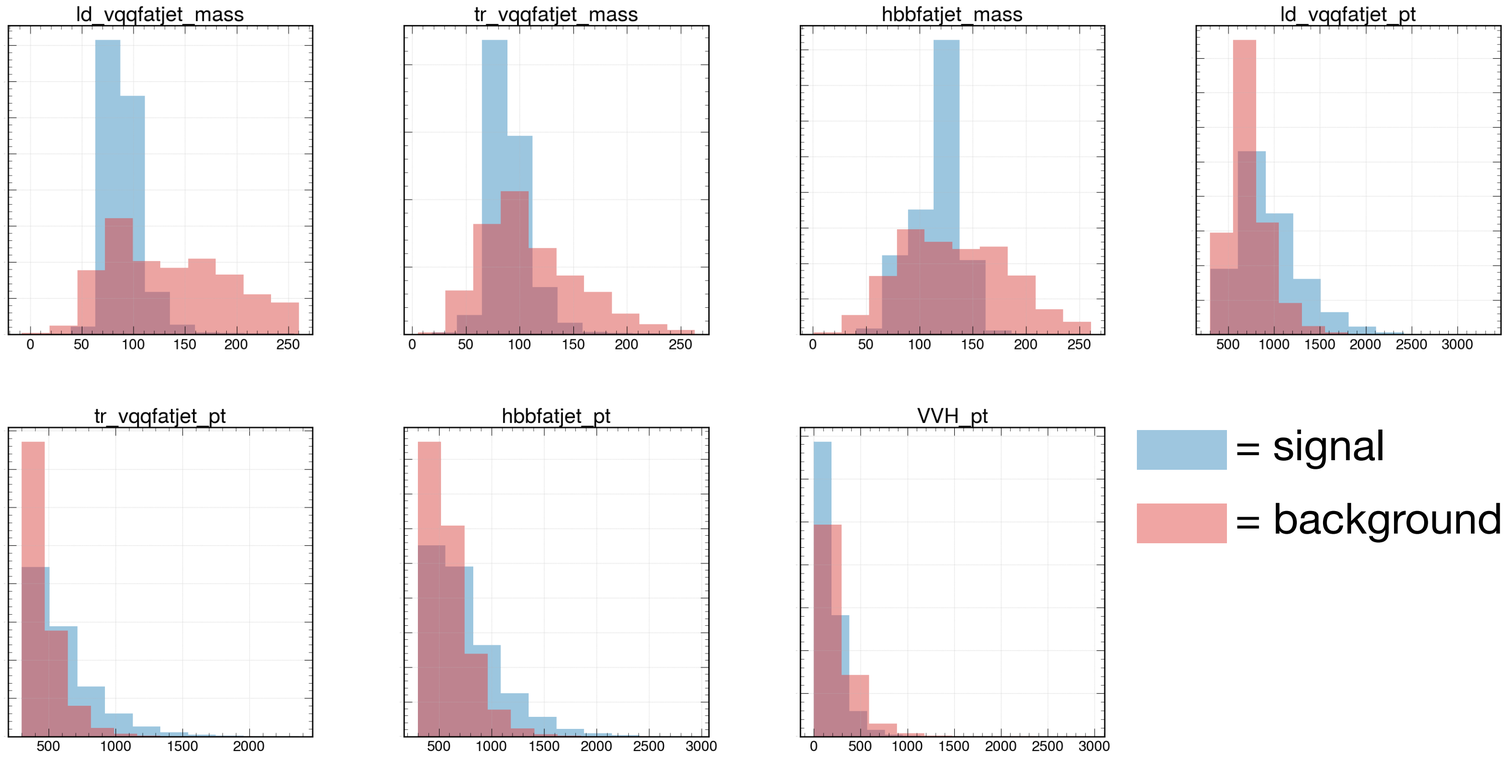
- Plotting sum of all VBS VVH signals here
- Selections applied: skim, HLT triggers, gen-level H, W, Z decay hadronically
- Practically 0 events with more than three fatjets



Cutflow

	QCD		TTHad		TT1L		TTW		TTH		SingleTop		Bosons		TotalBkg		VBSVVH	
cut	raw	wgt	raw	wgt	raw	wgt	raw	wgt	raw	wgt	raw	wgt	raw	wgt	raw	wgt	raw	wgt
Bookkeeping	23348898	137060539.91	11563843	748039.60	1952781	85549.20	291342	2574.92	1457254	1341.75	801492	53435.60	18001906	1512567.11	57417516	139464048.10	323711	175.15
SaveSystWeights	23348898	137060539.91	11563843	748039.60	1952781	85549.20	291342	2574.92	1457254	1341.75	801492	53435.60	18001906	1512567.11	57417516	139464048.10	323711	175.15
PassesEventFilters	23228634	136865407.91	11539080	746478.00	1946259	85267.20	290178	2566.97	1452921	1337.93	799328	53280.13	17946508	1508632.44	57202908	139262970.58	320310	173.33
PassesTriggers	19971141	88702387.12	8948283	574638.30	1610549	70244.26	261105	2236.23	1270751	1142.71	620348	41251.69	15012150	1120027.71	47694327	90511928.03	314509	168.32
SelectLeptons	19971141	88702387.12	8948283	574638.30	1610549	70244.26	261105	2236.23	1270751	1142.71	620348	41251.69	15012150	1120027.71	47694327	90511928.03	314509	168.32
NoLeptons	19971141	88702387.12	8948283	574638.30	1610549	70244.26	261105	2236.23	1270751	1142.71	620348	41251.69	15012150	1120027.71	47694327	90511928.03	314509	168.32
SelectFatJets	19971141	88702387.12	8948283	574638.30	1610549	70244.26	261105	2236.23	1270751	1142.71	620348	41251.69	15012150	1120027.71	47694327	90511928.03	314509	168.32
TriggerPlateauCuts	11032313	17507476.58	2402330	151403.21	577399	24859.33	117668	951.51	477636	414.98	178184	12210.73	4786171	331011.06	19571701	18028327.40	240732	129.81
Geq3FatJets	625880	395252.58	147659	9753.17	30742	1372.63	18012	110.44	50985	45.99	9319	873.52	211979	13105.95	1094576	420514.28	58085	31.64
AllMerged_ReplacePNetsQCD	625880	395252.58	147659	9753.17	30742	1372.63	18012	110.44	50985	45.99	9319	873.52	211979	13105.95	1094576	420514.28	58085	31.64
AllMerged_SelectVVHFatJets	625880	395252.58	147659	9753.17	30742	1372.63	18012	110.44	50985	45.99	9319	873.52	211979	13105.95	1094576	420514.28	58085	31.64
AllMerged_SetPtSortedFatJetVariables	625880	395252.58	147659	9753.17	30742	1372.63	18012	110.44	50985	45.99	9319	873.52	211979	13105.95	1094576	420514.28	58085	31.64
AllMerged_SelectJets	625880	393124.82	147659	9689.05	30742	1363.65	18012	109.84	50985	45.68	9319	868.00	211979	13042.54	1094576	418243.58	58085	31.24
AllMerged_SelectVBSJets	300296	158240.71	94313	6171.85	19389	854.93	11634	58.85	32458	30.18	5298	477.98	85560	5146.20	548948	170980.70	32951	17.71
AllMerged_SaveVariables	300296	158240.71	94313	6171.85	19389	854.93	11634	58.85	32458	30.18	5298	477.98	85560	5146.20	548948	170980.70	32951	17.71
AllMerged_MjjGt500	81337	36481.09	23578	1512.52	5146	223.48	2229	13.20	7651	7.00	1675	150.36	19854	1179.31	141470	39566.96	27420	14.71
AllMerged_detajjGt3	63871	31003.64	19862	1269.03	4331	187.52	1882	10.62	6538	5.94	1413	126.67	16215	959.46	114112	33562.87	27201	14.59
AllMerged_XbbGt0p9	3698	1872.43	6683	427.79	1700	73.28	688	3.72	3624	2.75	449	44.77	2074	83.07	18916	2507.80	18272	9.56
AllMerged_XVqqGt0p9	12	8.13	59	3.78	7	0.33	65	0.28	80	0.07	12	1.70	75	0.93	310	15.20	5717	3.09
AllMerged_STGt1300	12	8.13	54	3.45	6	0.28	61	0.27	71	0.06	12	1.70	74	0.86	290	14.75	5677	3.06
AllMerged_HbbMSDLt150	8	6.89	26	1.61	5	0.24	38	0.24	49	0.03	7	1.05	66	0.56	199	10.63	5567	3.01
AllMerged_VqqMSDLt120	3	5.56	13	0.71	2	0.08	19	0.04	13	0.01	3	0.47	44	0.30	97	7.18	5170	2.83

BDT



Using matplotlib's weird automatic binning (for visualization purposes only)

Parameter	Value	Description*
objective	binary:logistic	Learning objective; 'binary:logistic' specifies logistic regression for binary classification, output probability
eta	0.1	Step size shrinkage (alias: learning_rate)
max_depth	3	Max. depth of tree: larger = more complex = more prone to overfitting
verbosity	1	0 (silent), 1 (warning), 2 (info), 3 (debug)
nthread	8	Number of parallel threads
eval_metric	auc	Evaluation metrics for validation data. 'auc' = Area Under the Curve
subsample	0.6	Subsample ratio of the training instances
alpha	8.0	L1 regularization term on weights: Larger = more conservative
gamma	2.0	Min. loss rediction to make leaf (alias: min_split_loss)
lambda	1.0	L2 regularization term on weights: Larger = more conservative
min_child_weight	1.0	Minimum sum of instance weight (hessian) needed in a child
colsample_bytree	1.0	The subsample ratio of columns when constructing each tree
scale_pos_weight	2456.3	Control the balance of positive and negative weights, useful for unbalanced classes

*From: https://xgboost.readthedocs.io/en/stable/parameter.html

UC San Diego

Sanity Check

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_0)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg (batch size = 40K)

IV. APPLICATIONS

This section explores the efficacy of single and double DisCo in some applications of the ABCD method.

A. Simple example: Three-dimensional Gaussian random variables

We begin with a simple example to build some intuition and validate our methods. Consider a three-dimensional space (X_0, X_1, X_2), where the signal and background are both multivariate Gaussian distributions. We choose the means $\vec{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ for background and signal as

1

$$\vec{\mu}_b = (0,0,0), \qquad \boldsymbol{\Sigma}_b = \sigma_b^2 \begin{pmatrix} 1 & \rho_b & 0 \\ \rho_b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\sigma_b = 1.5, \qquad \rho_b = -0.8, \tag{4.1}$$

and

2

$$\vec{\mu}_s = (2.5, 2.5, 2), \qquad \boldsymbol{\Sigma}_s = \sigma_s^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \sigma_s = 1.5. \tag{4.2}$$

3

So for the background, all three features are centered at the origin and features X_0 and X_1 are correlated with each other but independent of X_2. For the signal, all three features are independent but are centered away from the origin. The first feature X_0 will play the role of the known feature for single DisCo in Sec. III.

All of the neural networks presented in this section use three hidden layers with 128 nodes per layer. The rectified linear unit (ReLU) activation function is used for the intermediate layers and the output is a sigmoid function. A hyperparameter of $\lambda = 1000$ is used for both single and double DisCo to ensure total decorrelation. The single DisCo training converged after 100 epochs while the double DisCo training required 200 epochs. Other networks only needed ten epochs. The double DisCo networks

4

were trained using a single neural network with a two-dimensional output. All models were trained using Tensorflow [89] through Keras [90] with Adam [91] for optimization. Two million examples were generated with 15% used for testing. A batch size of 1% of the total was used for all networks to ensure an accurate calculation of the DisCo term in the relevant loss functions.

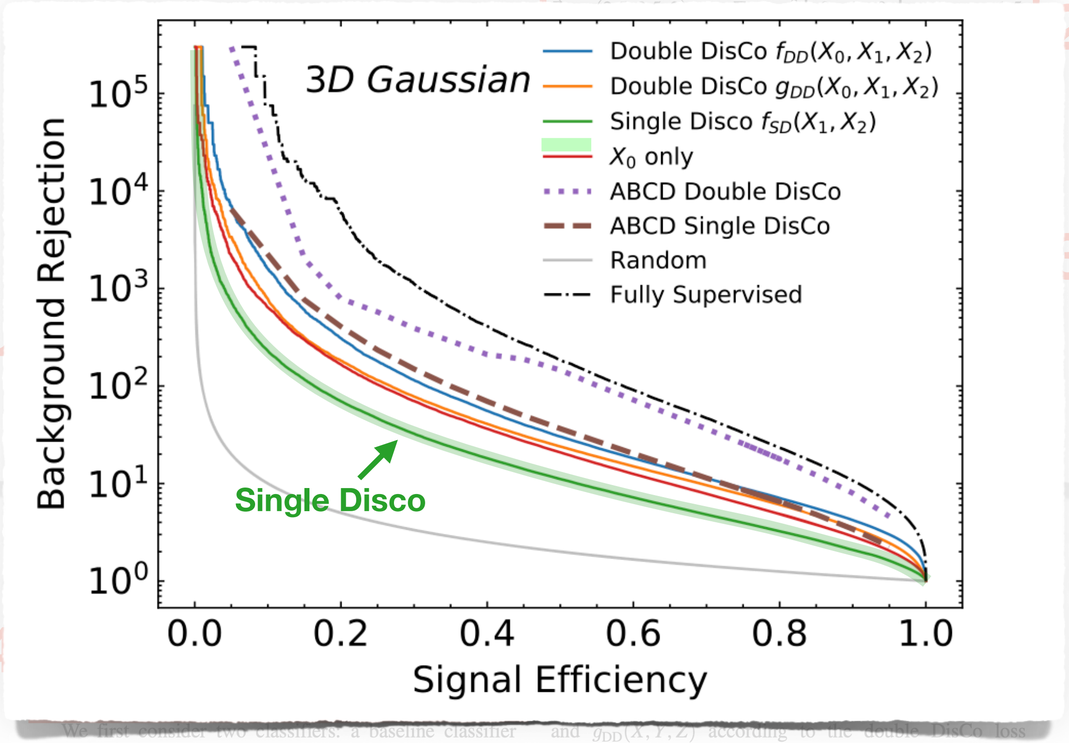
We first consider two classifiers: a baseline classifier $f_{\mathrm{BL}}(X_1, X_2)$ trained only on X_1 and X_2 and a single DisCo classifier $f_{\mathrm{SD}}(X_1, X_2)$ which includes a penalty for correlations between f_{SD} and X_0. The values of these classifiers for events drawn from the distributions are plotted in Fig. 3 against the X_0, X_1, or X_2 values of these events. We see that even though X_0 was not used in the training of the baseline, the classifier output is still correlated with X_0 because of the correlations between X_0 and X_1. In contrast to the baseline classifier, the single DisCo classifier is independent of both X_0 and X_1 and is simply a function of X_2. Intuitively, it makes sense that a classifier that must be independent of X_0 must also be independent of X_1. This is justified rigorously in Appendix B.

For double DisCo, we train two classifiers $f_{\mathrm{DD}}(X, Y, Z)$ and $g_{\mathrm{DD}}(X, Y, Z)$ according to the double DisCo loss function. The results are illustrated in Fig. 4. The first classifier depends mostly on Z and the second classifier depends mostly on X and Y. However, the residual dependence on all three observables is not a deficit of the training procedure: even though the three random variables are separable into two independent subsets (X, Y) and Z, the two classifiers learned by double DisCo

Sanity Check

- **Goal:** repeat the first example in the PRL paper (3D gaussian variables)
- **(1)** and **(2)** define the 3D gaussians
- **(3)** and **(4)** give the rest:
  - Input: X_1, X_2 (DisCo target: X_0)
  - NN architecture: 3 hidden layers; 128 nodes per layer; ReLU between layers; sigmoid output
  - λ = 1000, Adam optimizer
  - 2M sig, 2M bkg (batch size = 40K)

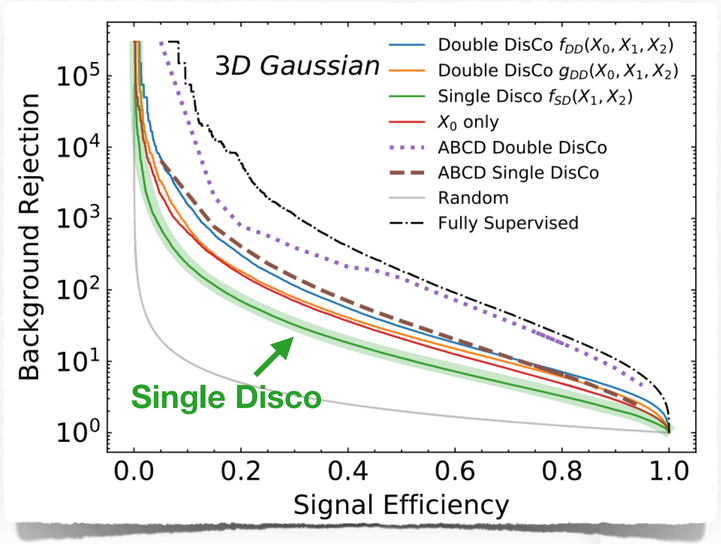
Target: recreate their plots, e.g.

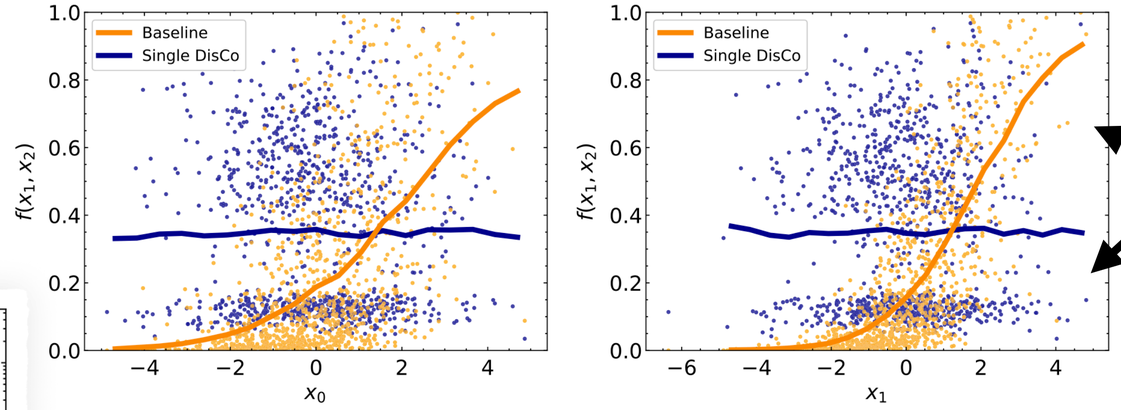


TPR = TP/P = (true positives)/(positives)
FPR = FP/N = (false positives)/(negatives)

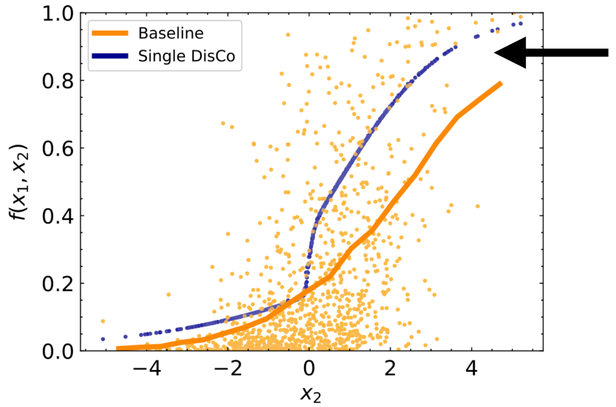
3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}_{y=0}(f_{SD}(X_1, X_2), X_0)$$





Two distinct clusters



Single DisCo is a perfect line

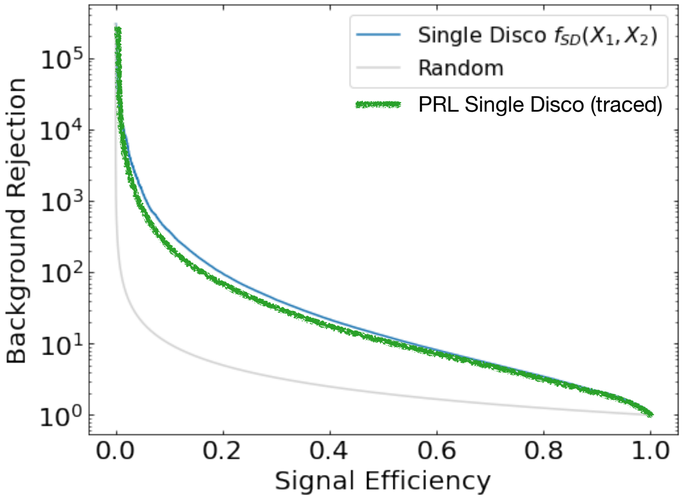
FIG. 3. Scatter plots showing the relationship (or lack thereof) between the three random variables X_0, X_1, and X_2 and (1) a baseline classifier $f_{\text{BL}}(X_1, X_2)$ trained on X_1 and X_2 with no regularization, and (2) a classifier $f_{\text{SD}}(X_1, X_2)$ trained with the single DisCo loss function that penalizes correlations with X_0. Only the background events are shown in these plots. The solid lines are the averages of the classifiers over events with the same value of X_0, X_1, or X_2. In the third panel, the scatter of the single DisCo classifier is already a line, so no average is needed.

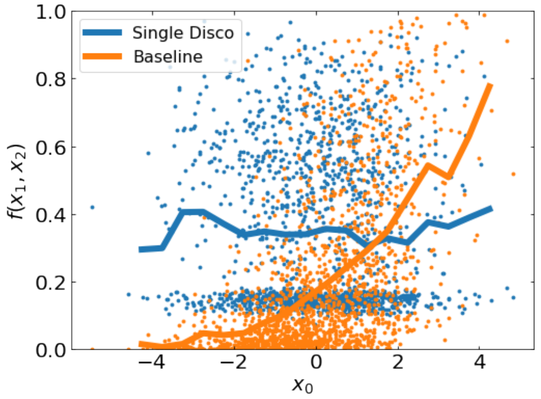
UC San Diego

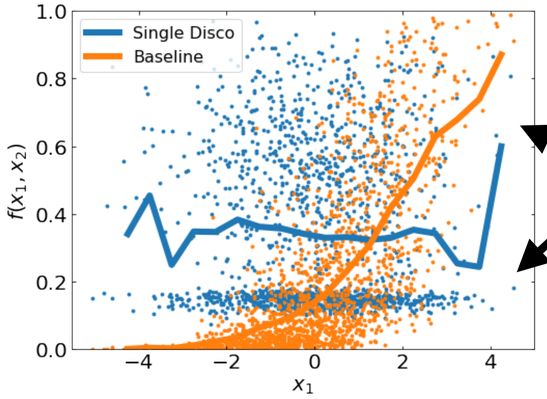
3D Gaussians: λ = 1000 DisCo

$$\mathscr{L} = \mathscr{L}_{BCE}(f_{SD}(X_1, X_2), y) + 1000 \times \text{dCorr}_{y=0}(f_{SD}(X_1, X_2), X_0)$$

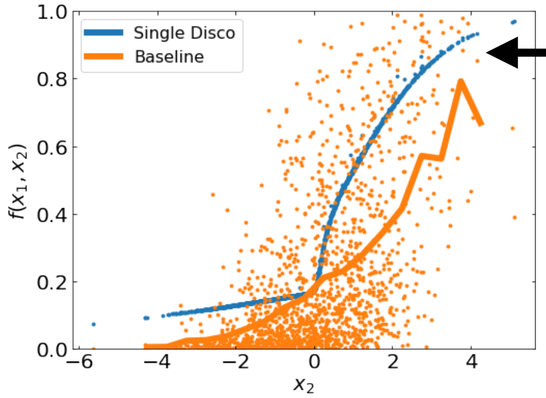
My ROC curve matches ✓







Two distinct clusters ✓



Single DisCo is a perfect line ✓

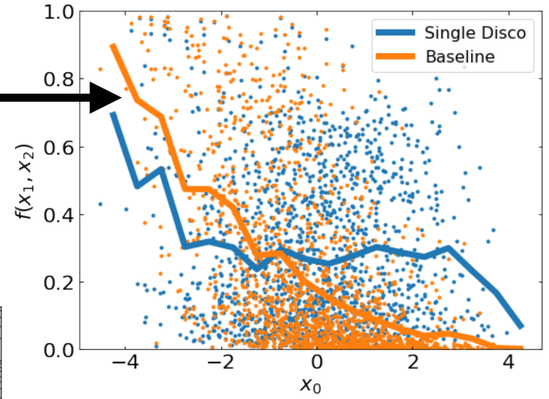
My plots match those in the PRL paper ✓

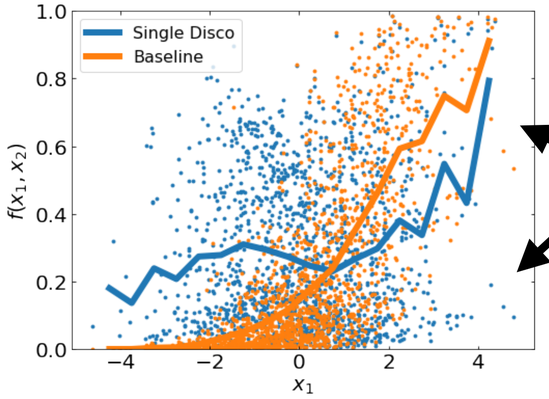
UC San Diego

3D Gaussians: λ = 1000 DisCo

Originally, we were unable to reproduce 3D Gaussian example

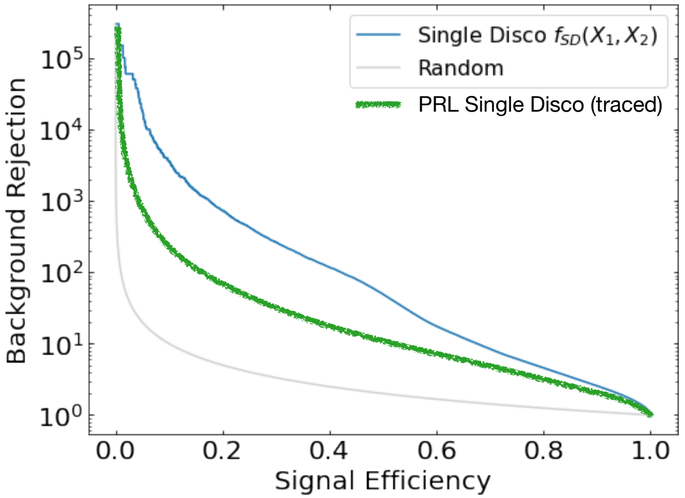
Baseline avg. opposite of that in the PRL paper

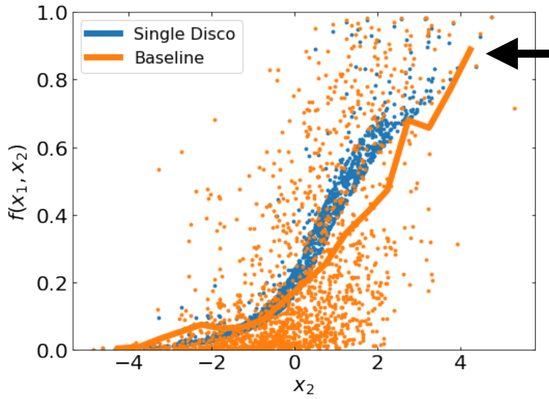




No distinct clusters

My ROC curve is different



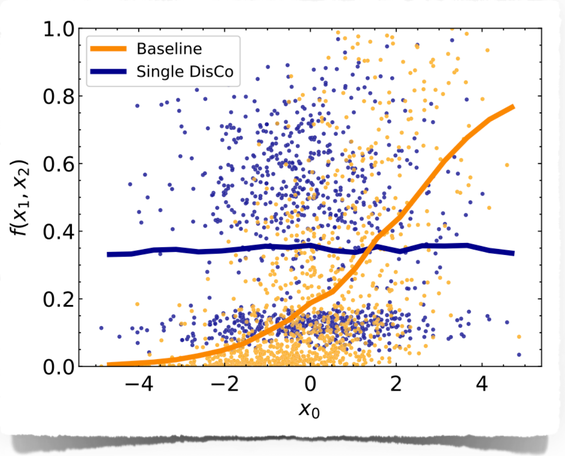


Single DisCo not a perfect line (more similar to Baseline avg.)

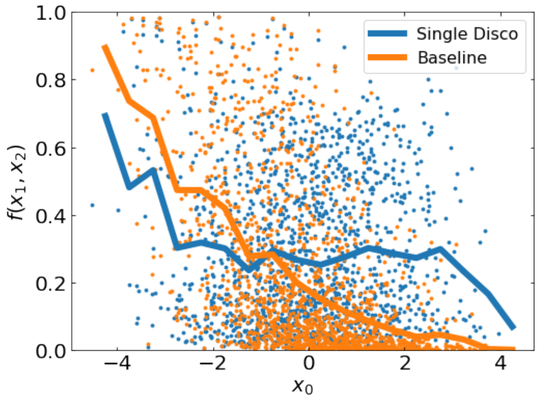
3D Gaussians: λ = 1000 DisCo

Spotted two issues (typos?) with PRL paper

Fig. 3



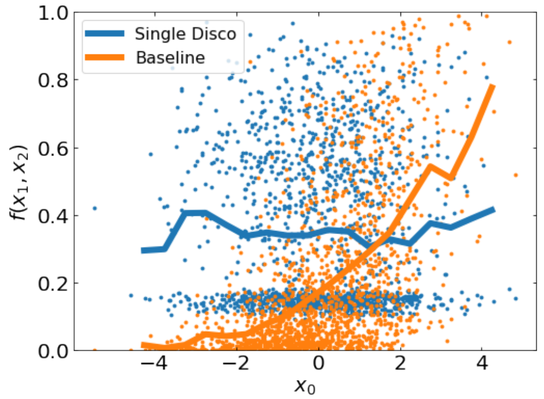
vs.



Does not match PRL

dCorr² → dCorr

Set ρ = +0.8



Matches PRL ✓

Eq. 3.1

$$\mathcal{L}[f(X)] = \mathcal{L}_{\text{classifier}}[f(X), y] + \lambda \text{dCorr}^2_{y=0}[f(X), X_0],$$

dCorr² → dCorr

$$\lambda \text{dCorr}_{y=0}[f(X), X_0]$$

Eq. 4.1

$$\vec{\mu}_b = (0,0,0), \qquad \Sigma_b = \sigma_b^2 \begin{pmatrix} 1 & \rho_b & 0 \\ \rho_b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.8 & 0 \\ -0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\sigma_b = 1.5, \qquad \rho_b = -0.8,$$

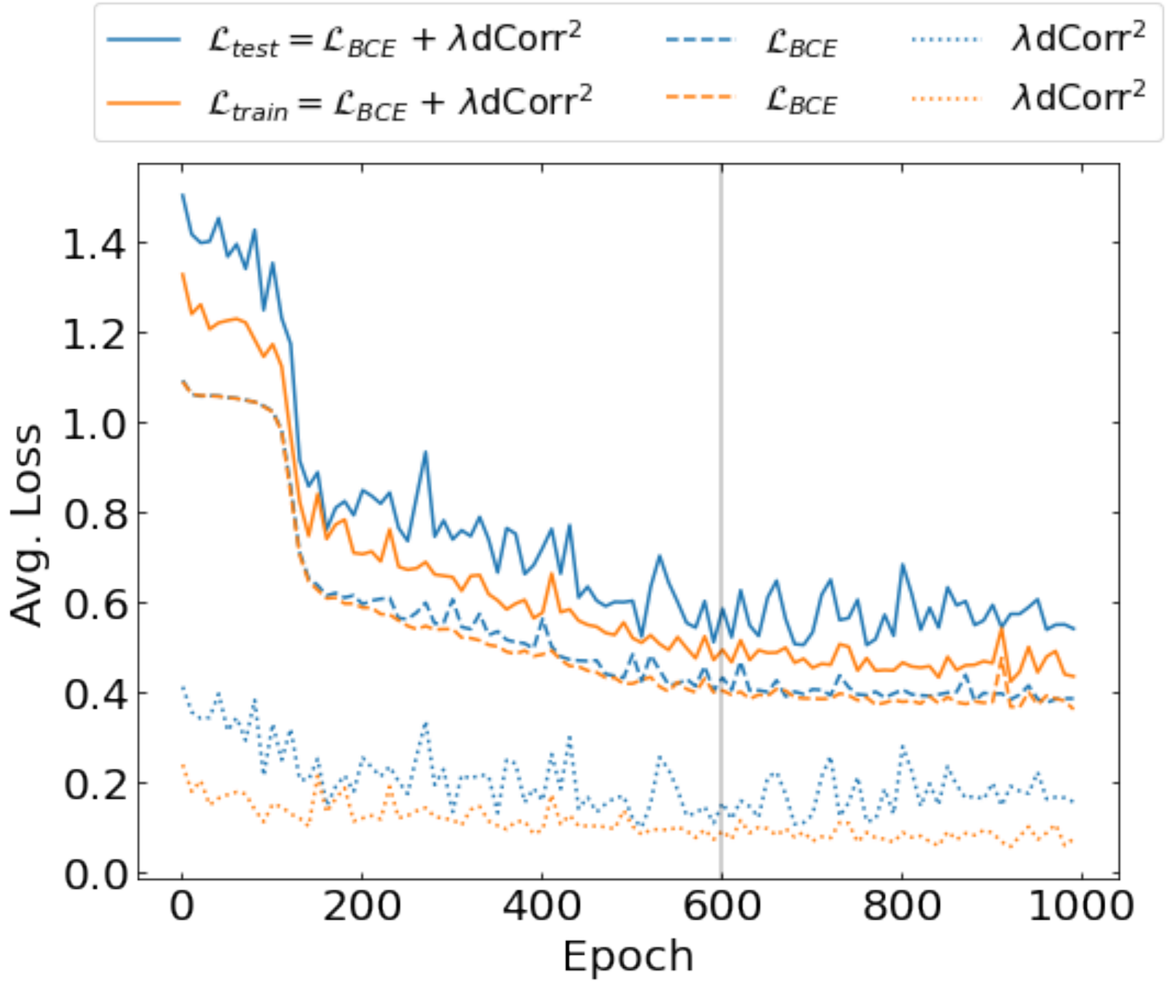
Set ρ = +0.8

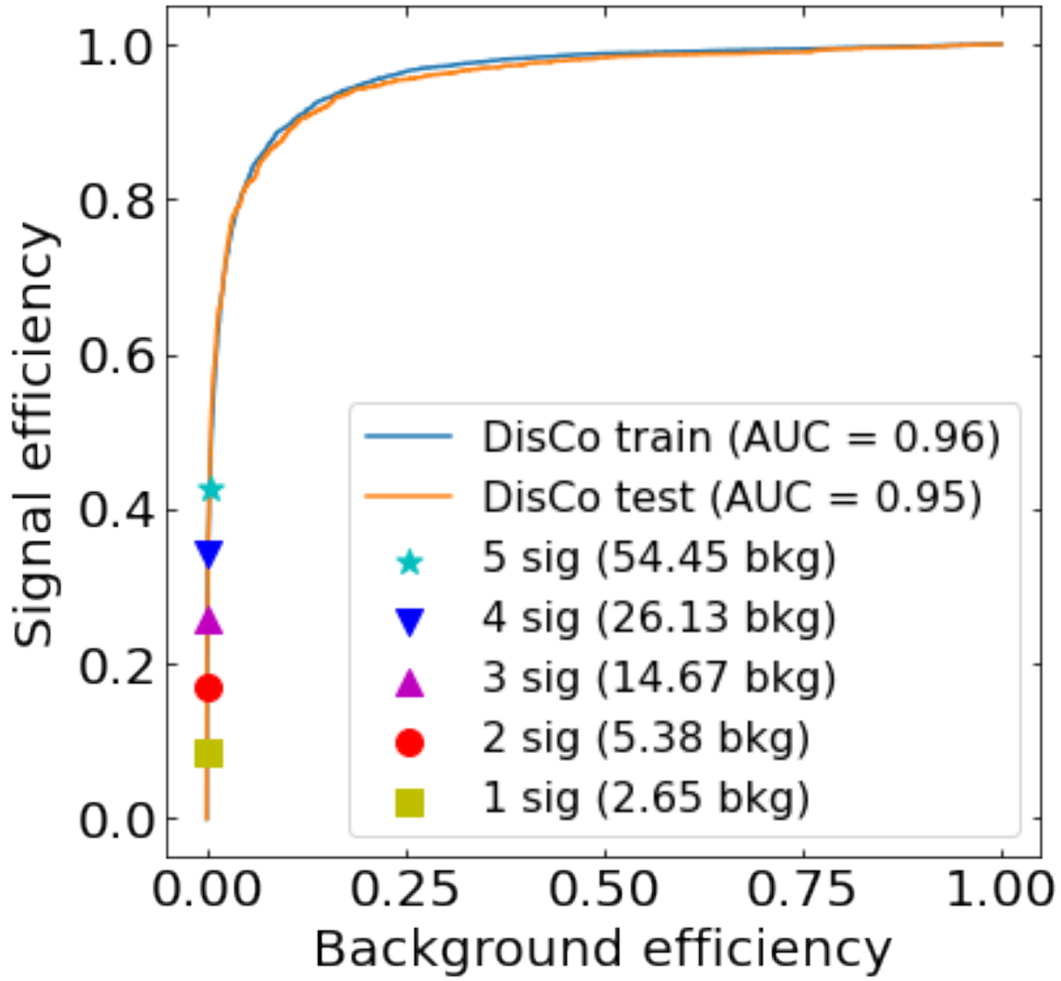
$$\Sigma_b = \sigma_b^2 \begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)

$$\mathscr{L} = \mathscr{L}_{BCE}(f(\vec{x}), y) + 30 \times \text{dCorr}_{y=0}(f(\vec{x}), |\Delta\eta_{jj}|)$$



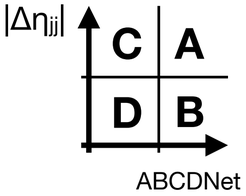


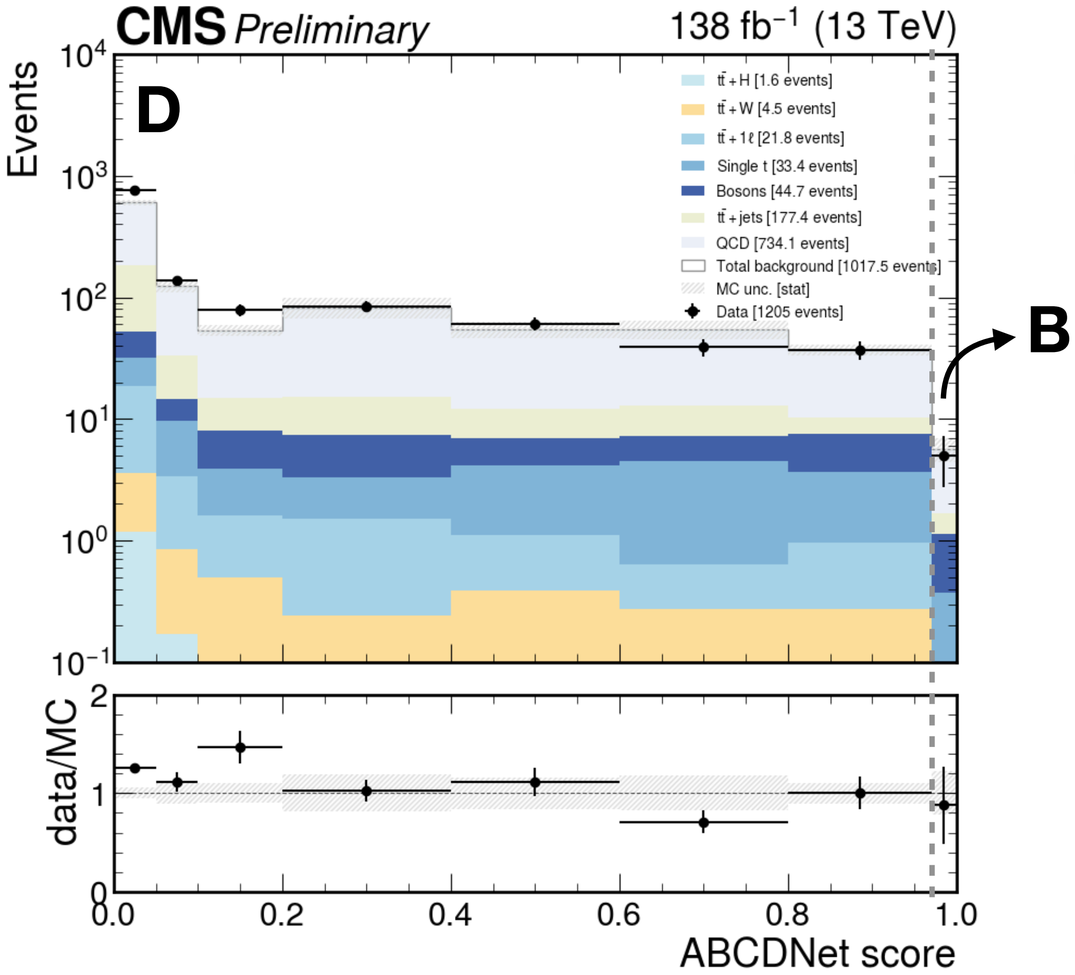
Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

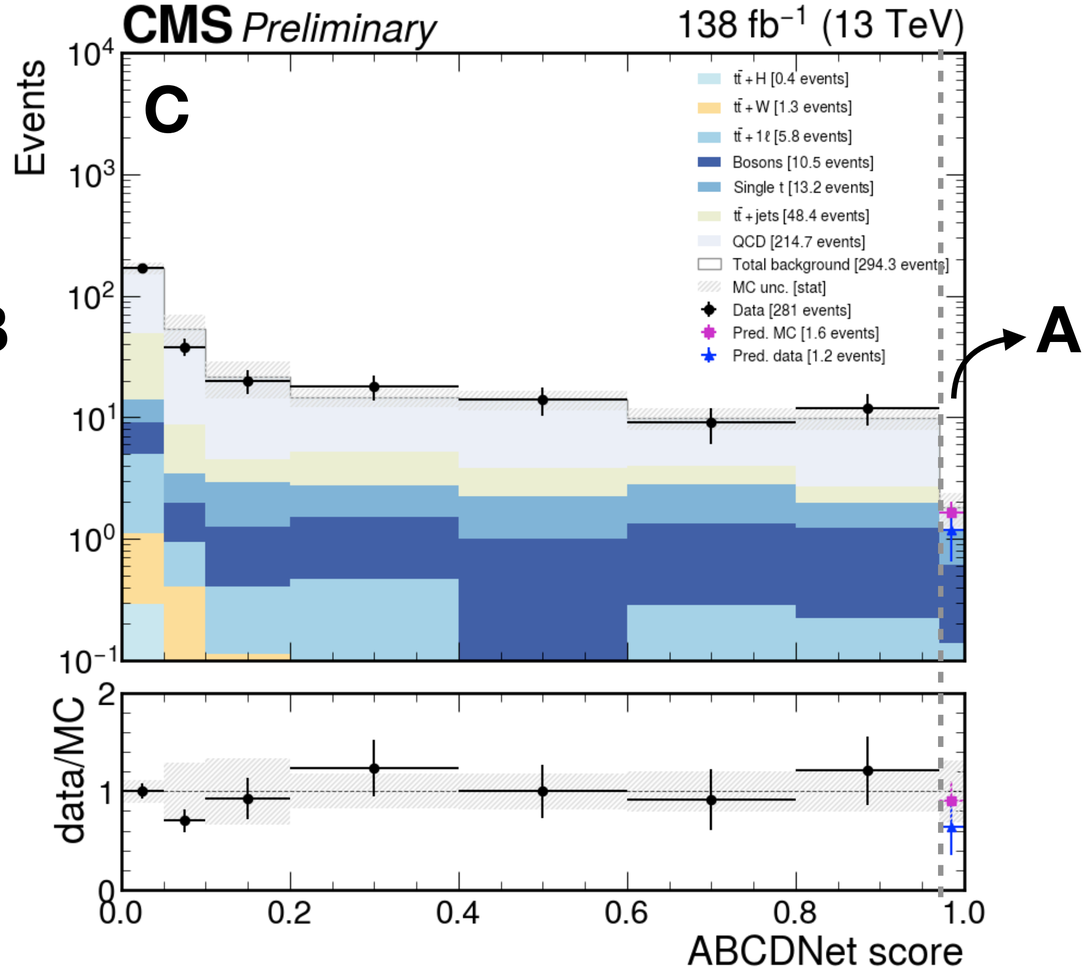
UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)

$$\mathscr{L} = \mathscr{L}_{BCE}(f(\vec{x}), y) + 30 \times \mathrm{dCorr}_{y=0}(f(\vec{x}), |\Delta\eta_{jj}|)$$



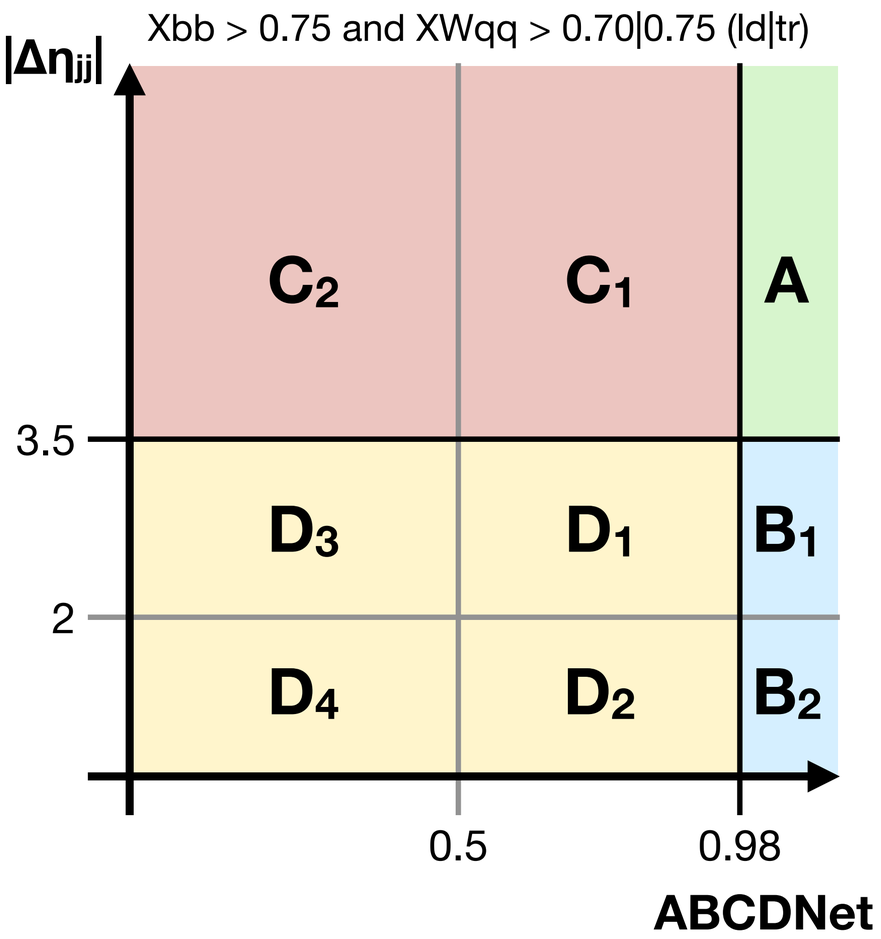




Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.79**	0.33	**3.00**	0.05	–	–
B	2.74	0.83	0.33	0.02	1	1.00
C	311.94	28.19	3.30	0.05	304	17.44
D	678.14	33.12	0.58	0.02	787	28.05

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.26 \pm 0.40 \text{ (MC)} \\ \mathbf{0.39 \pm 0.39} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	1.02	0.51	0.10	0.01	0	0.00
B_2	1.73	0.65	0.20	0.01	1	1.00
D_1	26.46	3.49	0.16	0.01	33	5.74
D_2	46.22	7.62	0.29	0.01	40	6.32

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 0.83 \pm 0.85 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	26.46	3.49	0.16	0.01	**33**	5.74
D_2	46.22	7.62	0.29	0.01	40	6.32
D_3	258.12	21.92	0.05	0.01	289	17.00
D_4	347.34	23.37	0.09	0.01	425	20.62

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{27.20 \pm 4.77} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	35.60	6.15	2.88	0.05	**30**	5.48
D_1	26.46	3.49	0.16	0.01	33	5.74
C_2	276.34	27.51	0.42	0.02	274	16.55
D_3	258.12	21.92	0.05	0.01	289	17.00

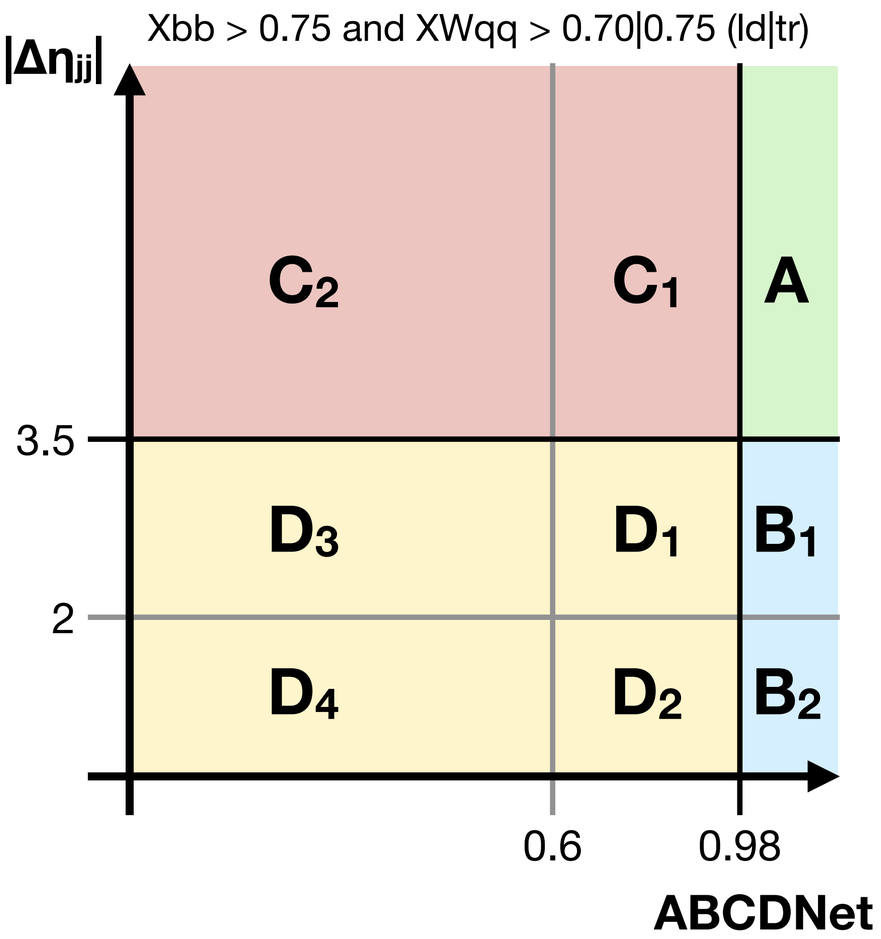
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{31.29 \pm 6.05} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.79**	0.33	**3.00**	0.05	–	–
B	2.74	0.83	0.33	0.02	1	1.00
C	311.94	28.19	3.30	0.05	304	17.44
D	678.14	33.12	0.58	0.02	787	28.05

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.26 \pm 0.40 \text{ (MC)} \\ \mathbf{0.39 \pm 0.39} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	1.02	0.51	0.10	0.01	0	0.00
B_2	1.73	0.65	0.17	0.01	1	1.00
D_1	19.55	2.93	0.15	0.01	20	4.47
D_2	36.22	7.30	0.27	0.01	30	5.48

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 0.67 \pm 0.69 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	19.55	2.93	0.15	0.01	**20**	4.47
D_2	36.22	7.30	0.27	0.01	30	5.48
D_3	265.02	22.00	0.05	0.01	302	17.38
D_4	357.34	23.47	0.10	0.01	435	20.86

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{20.83 \pm 4.11} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	22.07	2.95	2.75	0.05	**22**	4.69
D_1	19.55	2.93	0.15	0.01	20	4.47
C_2	289.87	28.03	0.54	0.02	282	16.79
D_3	265.02	22.00	0.05	0.01	302	17.38

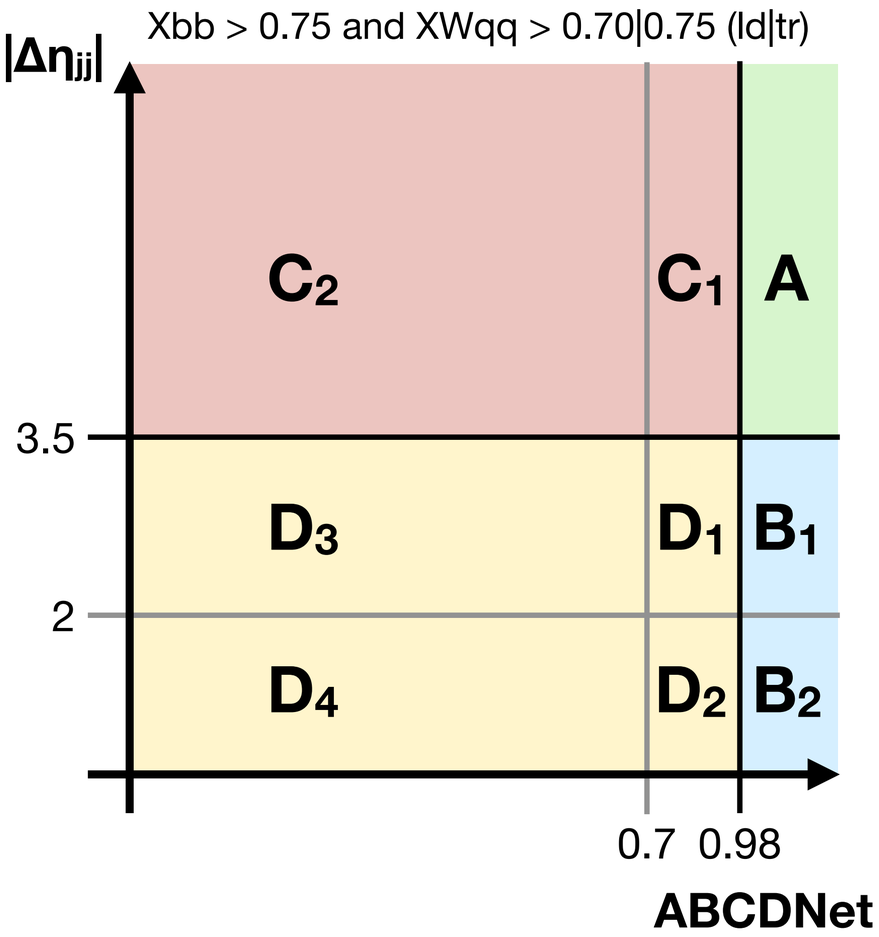
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{18.68 \pm 4.45} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.79**	0.33	**3.00**	0.05	–	–
B	2.74	0.83	0.33	0.02	1	1.00
C	311.94	28.19	3.30	0.05	304	17.44
D	678.14	33.12	0.58	0.02	787	28.05

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.26 \pm 0.40 \text{ (MC)} \\ \mathbf{0.39 \pm 0.39} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	1.02	0.51	0.10	0.01	0	0.00
B_2	1.73	0.65	0.13	0.01	1	1.00
D_1	14.64	2.52	0.14	0.01	15	3.87
D_2	27.41	7.06	0.24	0.01	22	4.69

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 0.68 \pm 0.72 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	14.64	2.52	0.14	0.01	**15**	3.87
D_2	27.41	7.06	0.24	0.01	22	4.69
D_3	269.94	22.05	0.06	0.01	307	17.52
D_4	366.16	23.55	0.13	0.01	443	21.05

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{15.25 \pm 3.44} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	17.54	2.66	2.57	0.04	**19**	4.36
D_1	14.64	2.52	0.14	0.01	15	3.87
C_2	294.40	28.06	0.73	0.02	285	16.88
D_3	269.94	22.05	0.06	0.01	307	17.52

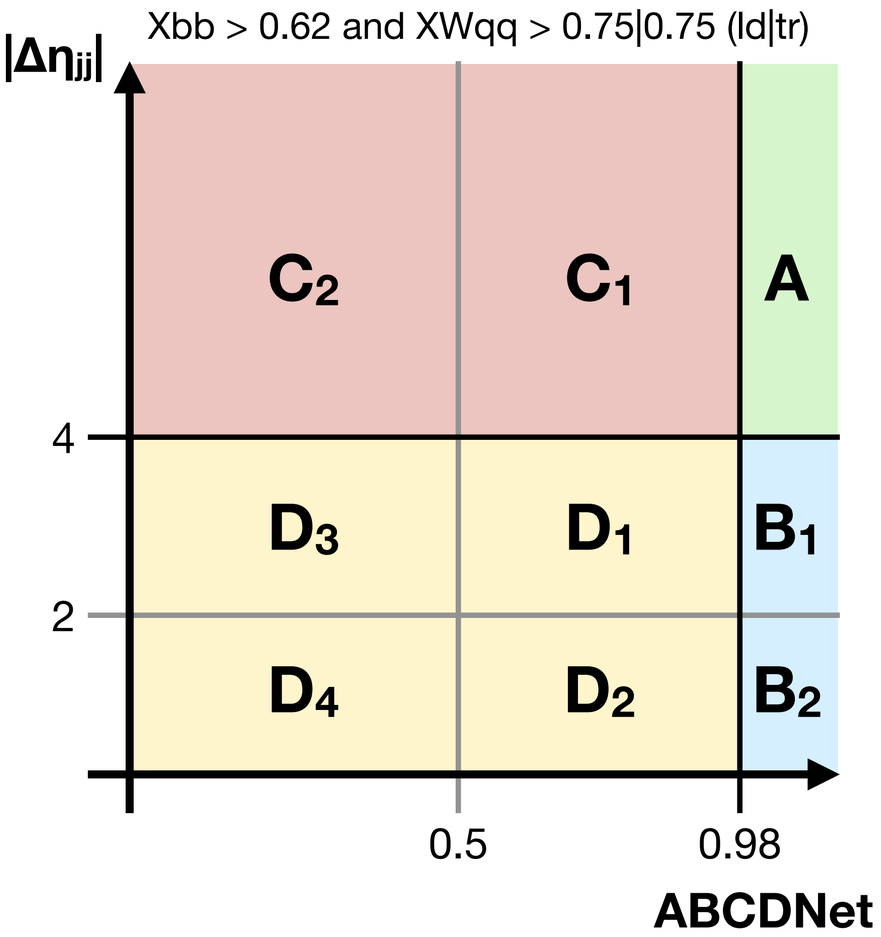
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{13.93 \pm 3.77} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.62**	0.55	**3.32**	0.05	–	–
B	4.31	1.12	0.45	0.02	3	1.73
C	246.35	25.17	2.86	0.05	216	14.70
D	797.78	36.54	0.60	0.02	947	30.77

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.33 \pm 0.38 \text{ (MC)} \\ \mathbf{0.68 \pm 0.40} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	2.50	0.94	0.18	0.01	2	1.41
B_2	1.80	0.61	0.27	0.01	1	1.00
D_1	37.31	4.11	0.19	0.01	46	6.78
D_2	48.50	7.70	0.26	0.01	40	6.32

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 1.15 \pm 1.18 \text{ (Data)} \checkmark$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	37.31	4.11	0.19	0.01	**46**	6.78
D_2	48.50	7.70	0.26	0.01	40	6.32
D_3	346.53	26.31	0.06	0.01	411	20.27
D_4	365.43	23.81	0.08	0.01	450	21.21

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{36.53 \pm 6.29} \text{ (Data)} \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	23.14	3.15	2.45	0.04	**20**	4.47
D_1	37.31	4.11	0.19	0.01	46	6.78
C_2	223.20	24.97	0.41	0.02	196	14.00
D_3	346.53	26.31	0.06	0.01	411	20.27

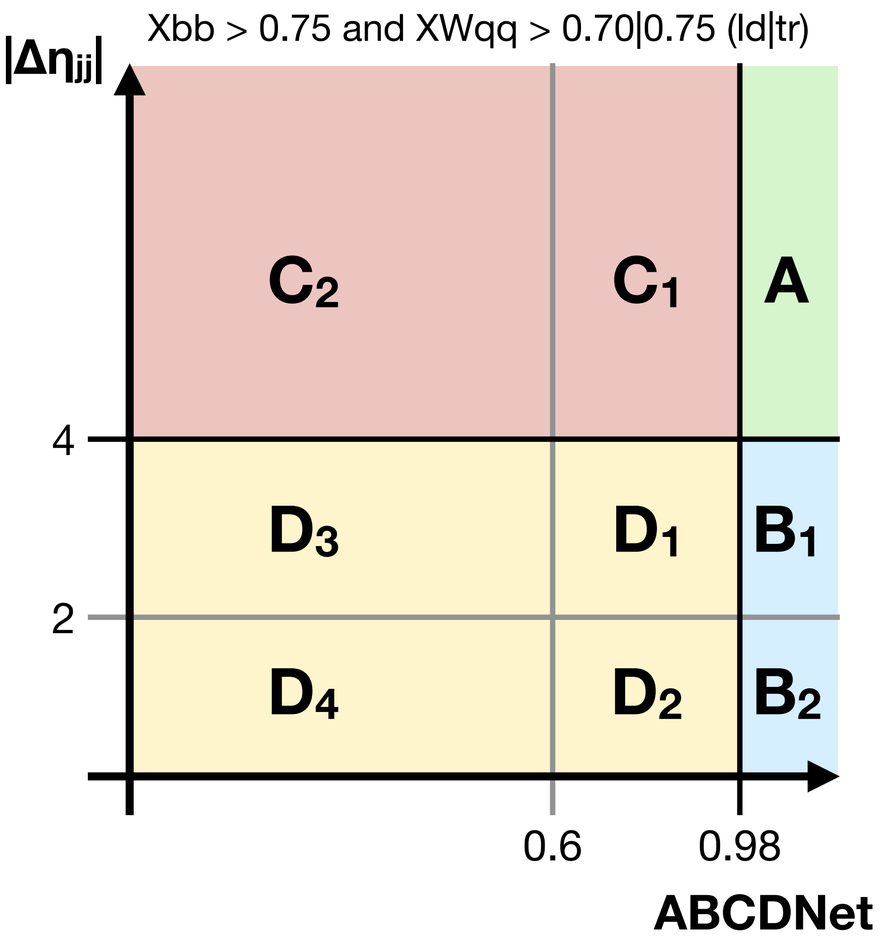
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{21.94 \pm 3.75} \text{ (Data)} \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.79**	0.33	**3.00**	0.05	–	–
B	2.74	0.83	0.33	0.02	1	1.00
C	311.94	28.19	3.30	0.05	304	17.44
D	678.14	33.12	0.58	0.02	787	28.05

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.26 \pm 0.40 \text{ (MC)} \\ \mathbf{0.39 \pm 0.39} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	1.02	0.51	0.10	0.01	0	0.00
B_2	1.73	0.65	0.01	0.01	1	1.00
D_1	19.55	2.93	0.15	0.01	20	4.47
D_2	36.22	7.30	0.27	0.01	30	5.48

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 0.67 \pm 0.69 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	19.55	2.93	0.15	0.01	**20**	4.47
D_2	36.22	7.30	0.27	0.01	30	5.48
D_3	265.02	22.00	0.05	0.01	302	17.38
D_4	357.34	23.47	0.10	0.01	435	20.86

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{20.83 \pm 4.11} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	22.07	2.95	2.75	0.05	**22**	4.69
D_1	19.55	2.93	0.15	0.01	20	4.47
C_2	289.87	28.03	0.54	0.02	282	16.79
D_3	265.02	22.00	0.05	0.01	302	17.38

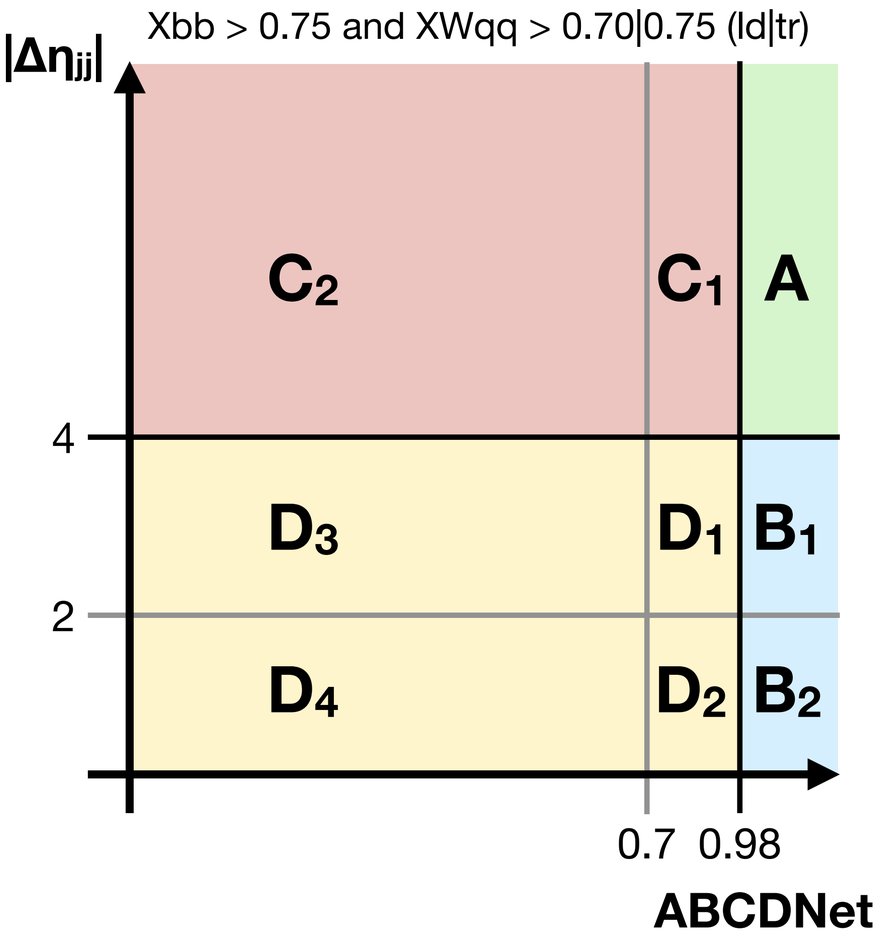
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{18.68 \pm 4.45} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 30 DisCo (Leaky ReLU)



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.79**	0.33	**3.00**	0.05	–	–
B	2.74	0.83	0.33	0.02	1	1.00
C	311.94	28.19	3.30	0.05	304	17.44
D	678.14	33.12	0.58	0.02	787	28.05

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 1.26 \pm 0.40 \text{ (MC)} \\ \mathbf{0.39 \pm 0.39} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	1.02	0.51	0.10	0.01	0	0.00
B_2	1.73	0.65	0.13	0.01	1	1.00
D_1	14.64	2.52	0.14	0.01	15	3.87
D_2	27.41	7.06	0.24	0.01	22	4.69

Low Stats.

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = 0.68 \pm 0.72 \text{ (Data) ?}$$

Low Stats.

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	14.64	2.52	0.14	0.01	**15**	3.87
D_2	27.41	7.06	0.24	0.01	22	4.69
D_3	269.94	22.05	0.06	0.01	307	17.52
D_4	366.16	23.55	0.13	0.01	443	21.05

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{15.25 \pm 3.44} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	17.54	2.66	2.57	0.04	**19**	4.36
D_1	14.64	2.52	0.14	0.01	15	3.87
C_2	294.40	28.06	0.73	0.02	285	16.88
D_3	269.94	22.05	0.06	0.01	307	17.52

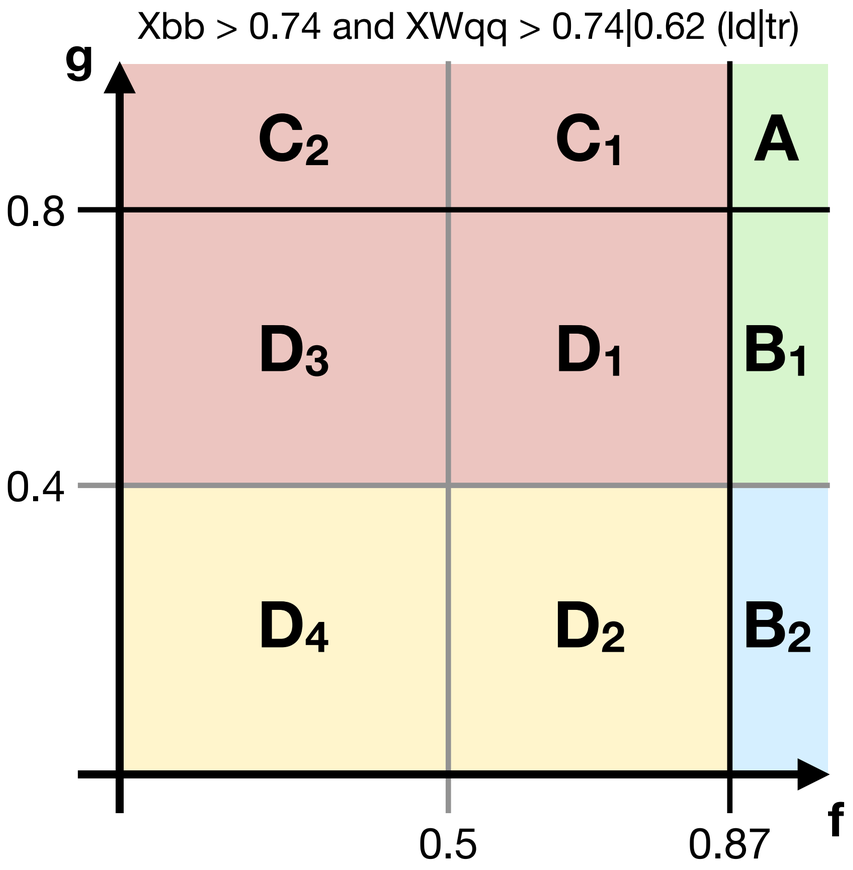
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{13.93 \pm 3.77} \text{ (Data) } \checkmark$$

ABCD works well with data in sidebands ⇒ method is valid!

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

ABCDNet: λ = 20 Double DisCo



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**1.38**	0.39	**4.52**	0.06	–	–
B	17.46	2.63	1.17	0.03	13	3.61
C	42.53	7.44	0.75	0.02	58	7.62
D	1273.13	49.10	1.41	0.03	1429	37.80

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 0.58 \pm 0.14 \text{ (MC)} \\ \mathbf{0.53 \pm 0.16} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
B_1	2.85	0.67	0.66	0.02	3	1.73
B_2	14.61	2.55	0.51	0.02	10	3.16
D_1	18.30	6.94	0.39	0.02	14	3.74
D_2	112.96	11.64	0.63	0.02	129	11.36

$$B_1^{pred} = B_2 \times \frac{D_1}{D_2} = \mathbf{1.09 \pm 0.46} \text{ (Data) ?}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
D_1	18.30	6.94	0.39	0.02	14	3.74
D_2	112.96	11.64	0.63	0.02	129	11.36
D_3	102.61	13.37	0.12	0.01	126	11.22
D_4	1039.27	45.26	0.27	0.01	1160	34.06

$$D_1^{pred} = D_2 \times \frac{D_3}{D_4} = \mathbf{14.01 \pm 1.80} \text{ (Data) } \checkmark$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	4.12	1.09	0.61	0.02	3	1.73
D_1	18.30	6.94	0.39	0.02	14	3.74
C_2	38.40	7.36	0.14	0.01	55	7.42
D_3	102.61	13.37	0.12	0.01	126	11.22

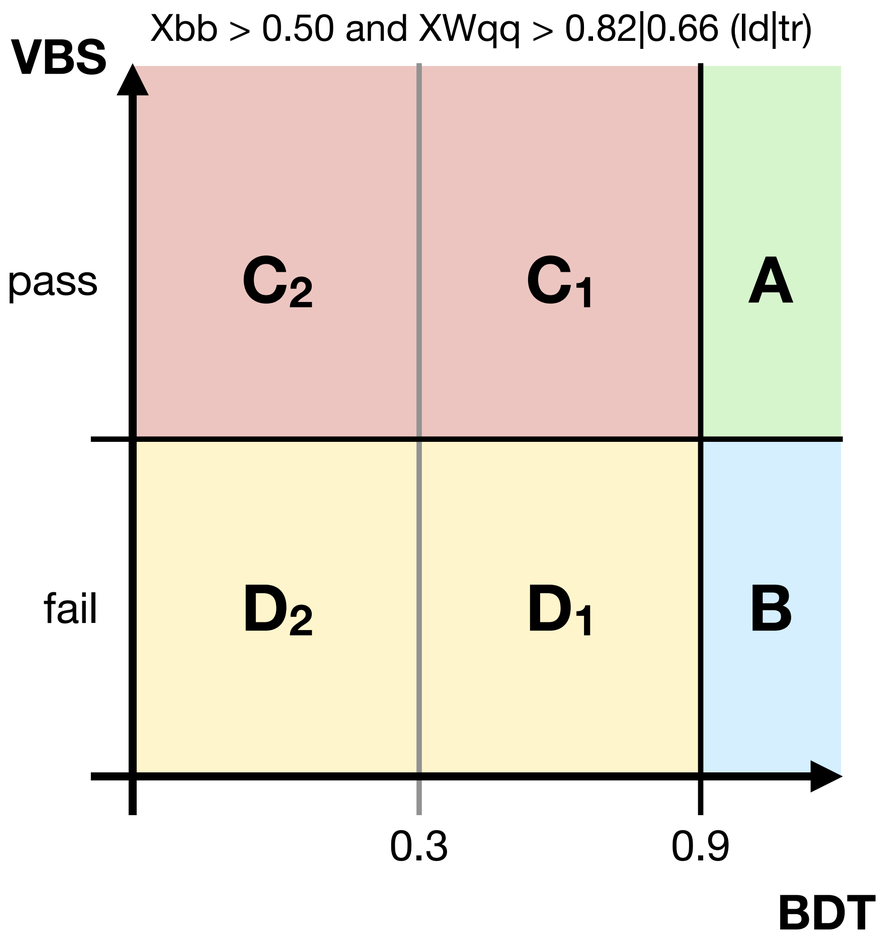
$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{6.11 \pm 1.91} \text{ (Data) } \checkmark$$

ABCD works OK in sidebands ⇒ method maybe valid?

Epoch = 600 | LR = 0.001 (constant) | λ = 30 | QCD norm | All features normalized

UC San Diego

BDT



Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.81**	0.28	**5.05**	0.06	–	–
B	20.87	2.62	1.03	0.03	25	5.00
C	175.92	20.87	1.15	0.03	172	13.11
D	1070.0	44.43	0.32	0.02	1190	34.50

$$A^{pred} = B \times \frac{C}{D} = \begin{bmatrix} 3.43 \pm 0.61 \text{ (MC)} \\ \mathbf{3.61 \pm 0.78} \text{ (Data)} \end{bmatrix}$$

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
A	**0.81**	0.28	**5.05**	0.06	–	–
B	20.87	2.62	1.03	0.03	25	5.00
C_1	19.72	7.16	1.01	0.03	19	4.36
D_1	136.7	13.42	0.25	0.01	142	11.92

$$A^{pred} = B \times \frac{C_1}{D_1} = \begin{bmatrix} 3.01 \pm 1.19 \text{ (MC)} \\ \mathbf{3.35 \pm 1.06} \text{ (Data)} \end{bmatrix}$$

✓

Region	Bkg	Bkg err	Sig	Sig err	Data	Data err
C_1	19.72	7.16	1.01	0.03	**19**	4.36
D_1	136.69	13.42	0.25	0.01	142	11.92
C_2	156.20	19.61	0.13	0.01	153	12.37
D_2	933.30	42.35	0.07	0.01	1048	32.37

$$C_1^{pred} = D_1 \times \frac{C_2}{D_3} = \mathbf{20.73 \pm 2.50} \text{ (Data)}$$ ✓

ABCD works well in sidebands, but predicted S/√B is worse than ABCDNet

UC San Diego